

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

3 Giovanni Laudanno¹, Richèl J.C. Bilderbeek¹, and Rampal S.
4 Etienne¹

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

7 September 27, 2018

8 **Abstract**

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

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

19 Currently implemented tree priors assume that speciation events are
20 independent, where we know that speciation can coincide, for example,
21 when trigger by a larger geographic change.

Here we explore the effects of ignoring speciation co-occurrence 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 &

48 Buckley 2004), clock model (Baele *et al.* 2012) or tree prior (Möller *et al.* 2018;
49 Yang & Rannala 2005) is known to affect the posterior.

50 Current phylogenetic tools assume that only a single speciation event can
51 occur at the same time. While this assumption is useful to construct a wide
52 variety of successful models [RJC: @gio: citation here] [GL: @richel:
53 basically all the models we know are based on this assumption: DDD,
54 PBD, BISSE, MUSSE, SECSSE, any other SSE, DAISIE etc. etc.
55 It's a very very general feature. Maybe being specific could lead the
56 reader to consider things that are, in the end, not essential to the
57 story we want to tell here. Do you still think we need it?], [RJC:
58 Yes, I think here would be a fine spot to cite some of those models,
59 I think BiSSE, DAISIE, DDD and PBD would be appropriate] they
60 disallow for environmental changes that trigger speciation on a large scale, for
61 example, the cichlid fish diversification in the African Great Lakes: Malawi,
62 Tanganyika and Victoria [RJC: @gio: citation here].

63 The (constant-rate) birth-death (BD) model embodies the common assump-
64 tion that only a single speciation event can occur at the same time. The MBD
65 model relaxes this assumption, allowing events in which large-scale environmen-
66 tal changes lead to a great number of species in relatively short time intervals.

67 [GL: If I described the process in the same way you report in the
68 example I would probably end up writing the same things that we
69 say a few lines below, where we describe the parameters. Don't you
70 think?] [RJC: You described the model in the Methods. I moved it
71 to here] In the MBD model, parameters λ and μ correspond, respectively, to
72 the usual per-species speciation and extinction rates. Additionally, ν is the rate
73 at which an environmental change is triggered. When that event is triggered,
74 all species at that moment have a probability q to speciate (independent on

75 λ). The number of species that speciate due to this can also be zero. [RJC:
76 Is this correct?].

77 Unfortunately, a tree prior according to this model, providing the probability
78 of a species tree under the MBD model, is unavailable in current Bayesian
79 phylogenetic tools. Whilst a likelihood equation has been derived ([RJC:
80 cite yourself here]), it has not been implemented as tree prior yet. There
81 are various reasons for this. First, the computation of the MBD likelihood
82 involves solving a set of non-linear differential equations [GL: @richel: are
83 they actually non-linear?], and while this computation is quite fast, it still
84 takes much more time than the corresponding probability of the BD model
85 which is a simple analytical formula. In a Bayesian MCMC chain, the tree prior
86 probability must be calculated many times, and hence the total computation
87 will take considerably longer with a PBD tree prior.

88 Here we aim to explore the effect of using the BD prior on MBD simulated
89 phylogenies. In brief, we simulate phylogenies with co-occurring speciation events
90 using the MBD process. Given this species tree, we simulate a DNA sequence
91 alignment. Then, we use BEAST2 on these alignments to infer a posterior of
92 phylogenies, using a BD prior. We quantify the difference between the (BD)
93 posterior phylogenies and the simulated (MBD) species tree. Furthermore, while
94 we evidently know the clock and site models used in the simulation, using a
95 different clock and/or site model prior in inference may compensate or increase
96 this difference between inferred and simulated tree. To study this, we also
97 explore the effect of a different clock and site model prior in inference.

98 The MBD model has 4 parameters, depicted in table 2. We pick values of ν
99 in such a way we expect a multiple speciation event to be triggered zero ($\nu = 0$),
100 once, twice, four and eight times [RJC: I assume you can calculate the
101 correct ν]. We pick values of q that are 0.0 (a speciation barrier at the triggered

event), 0.25, 0.5 and 1.0. We set our extinction rate μ to 0.1 in all simulation. As we select our phylogenies on their number of lineages, we calculate λ in a such a way that the mean expected number of lineages equals the desired numbers of taxa of 50, 100 and 200. For $\nu = 0$, the model falls back to a standard BD model. Note that the λ and q have different units and it is a misconception to think that for $\lambda = q$ (already impossible due to their units) the MBD model would reduce to a BD model.

We simulate protracted birth-death trees, using the MBD package (Laudanno 2018) 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, ν and q , we generate species trees with a crown age of 15 million years [GL: In general [15 million years] is ok for me. Keep in mind, though, that allowing multiple speciations may lead to an explosion in the number of species. Increasing the time by a factor of n usually means increasing the expected number of species at the present by a factor proportional to e^n] [RJCB: I know]. Only trees with the desired number of good taxa are kept.

From an (MBD) species tree, we create a BEAST2 posterior using the 'pirouette' (Bilderbeek 2018) R package: 'pirouette' first simulates a DNA alignment that has the same history as the species tree, using the **phangorn** package (Schliep 2011). 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 AAACC-CGGGTTT). Throughout evolutionary time, we use equal mutation rates between the four DNA nucleotides, also called the Jukes-Cantor (Jukes *et al.* 1969) nucleotide substitution model. The neat separation of the nucleotides is for visualization and debugging purposes and has no effect in any other way. The

129 equal amount of nucleotides does matter, assuring any nucleotide mutation is
130 equally likely to be observed.

131 In our Bayesian inference (see below) we use the same site model as the
132 (obviously correct) site model prior, but we also explore the effect of assuming a
133 more complex site model prior. We predict with the more complex substitution
134 model, that there will be more noise and hence our inference error will increase.
135 On the other hand, we dare not rule out that the inference error will decrease,
136 due to more flexibility in the more complex prior. We set the mutation rate in
137 such a way to maximize the information contained in the alignment. To do so,
138 we set the mutation rate such that we expect on average one (possibly silent)
139 mutation per nucleotide between crown age and present, which equates to $\frac{1}{15}$
140 mutations per million years. The DNA sequence length is chosen to provide a
141 resolution of 10^3 years, that is, to have one expected nucleotide change per 10^3
142 years per lineage on average. As one nucleotide is expected to have on average
143 one (possibly silent) mutation per 15 million years, $15 \cdot 10^3$ nucleotides result
144 in 1 mutation per alignment per 10^3 years (which is coincidentally the same
145 as Möller *et al.* 2018). The simulation of these DNA alignments follows a strict
146 clock model, which we will specify as one of the two clock models assumed in
147 the Bayesian inference (see below).

148 From here, the 'babette' R package (Bilderbeek & Etienne 2018) takes over
149 and converts the DNA alignment to a BEAST2 posterior. We set up the
150 BEAST2 analysis to assume either a Jukes-Cantor or GTR nucleotide sub-
151 stitution model. The Jukes-Cantor model is the correct one, as it is used for
152 simulating that alignment, where the GTR model is the site model that is picked
153 as a default by most users. For our clock model, we assume either a strict or
154 relaxed log-normal clock model. Also here, the strict clock model is the correct
155 one, as it is used for simulating the alignment, but the relaxed log-normal clock

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

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

176 The input trees generated with a $\nu = 0$, in which all BEAST2's assumptions
 177 are met, allow us to measure the noise of the experiment.

178 We produce one data set as a comma-separated file. The general data set
 179 has 144 [RJC*B*: recal*c*] different combinations of biological parameter com-
 180 binations, site and clock models. The data set to investigate sampling has 552
 181 [RJC*B*: recal*c*] different combinations of biological parameter combinations,
 182 site models, clock models and sampling methods. The experiment is compu-

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

Table 1: Glossary

183 tationally intensive: pilot experiments show that the experiment takes roughly
184 100 days of CPU time and 20 days of wall clock time (which includes the queued
185 waiting for computational resources) per replicate. Due to this, we choose to
186 perform ten replicates, so that the complete experiment will take an acceptable
187 time of roughly seven months.

188 We display the data set as an nLTT statistics distribution per parameter
189 combination as a violin plot. We only show the nLTT distributions that were
190 generated under the (correct) assumptions of a Jukes-Cantor site model and a
191 strict clock model, separated per sampling method used. We display the nLTT
192 statistic distributions separated per site or clock model in the supplementary
193 information.

194 2 Results

195 3 Glossary

196 [GL: bibliography is missing. The only bib file present does not cor-
197 respond to the bibliography showed in the pdf file.]

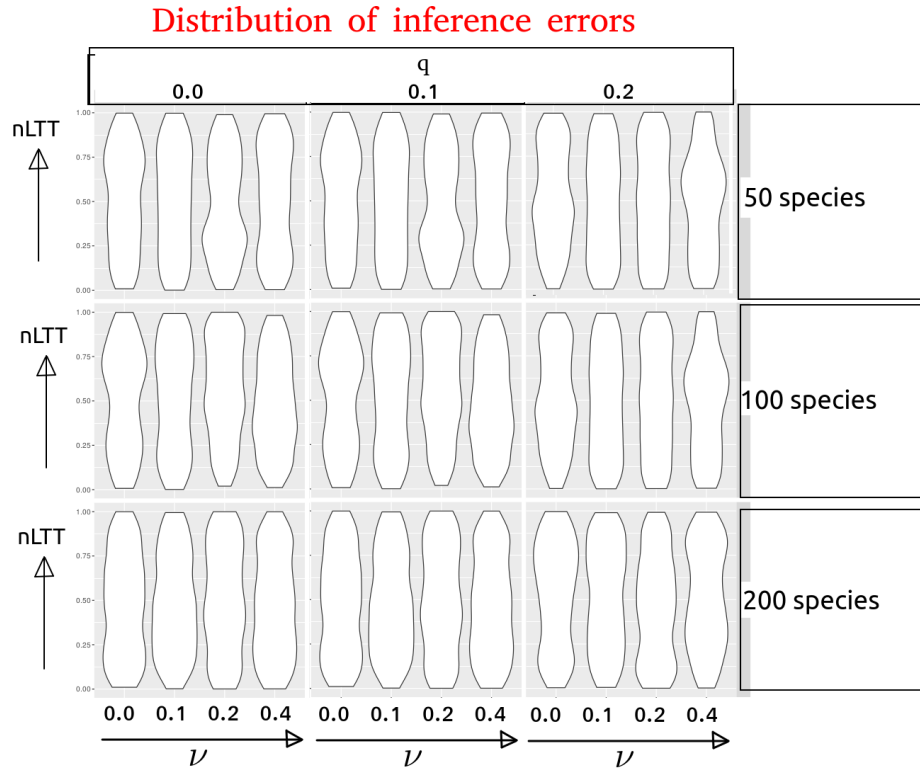


Figure 1: nLTT statistic distribution per biological parameter set, using the general data set, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

198 References

- 199 Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A. & Lemey, P. (2012) Ac-
 200 curate model selection of relaxed molecular clocks in bayesian phylogenetics.
 201 *Molecular biology and evolution*, **30**, 239–243.
- 202 Bilderbeek, R.J. (2018) *pirouette: create a posterior from a phylogeny*.
- 203 Bilderbeek, R.J. & Etienne, R.S. (2018) babette: Beauti 2, beast 2 and tracer
 204 for r. *Methods in Ecology and Evolution*.
- 205 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard,
 206 M.A., Rambaut, A. & Drummond, A.J. (2014) Beast 2: a software platform
 207 for bayesian evolutionary analysis. *PLoS computational biology*, **10**, e1003537.
- 208 Drummond, A.J. & Bouckaert, R.R. (2015) *Bayesian evolutionary analysis with*
 209 *BEAST*. Cambridge University Press.
- 210 Drummond, A.J. & Rambaut, A. (2007) Beast: Bayesian evolutionary analysis
 211 by sampling trees. *BMC evolutionary biology*, **7**, 214.
- 212 Felsenstein, J. (1981) Evolutionary trees from dna sequences: a maximum like-
 213 lihood approach. *Journal of molecular evolution*, **17**, 368–376.
- 214 Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R.,
 215 Huelsenbeck, J.P. & Ronquist, F. (2016) Revbayes: Bayesian phylogenetic
 216 inference using graphical models and an interactive model-specification lan-
 217 guage. *Systematic biology*, **65**, 726–736.
- 218 Huelsenbeck, J.P. & Ronquist, F. (2001) Mrbayes: Bayesian inference of phylo-
 219 genetic trees. *Bioinformatics*, **17**, 754–755.
- 220 Janzen, T. (2015) *nLTT: Calculate the NLTT Statistic*. R package version 1.1.

221 Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian compu-
222 tation of diversification rates from molecular phylogenies: introducing a new
223 efficient summary statistic, the nlrt. *Methods in Ecology and Evolution*, **6**,
224 566–575.

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

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

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

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

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

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

240 Schliep, K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**,
241 592–593.

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

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

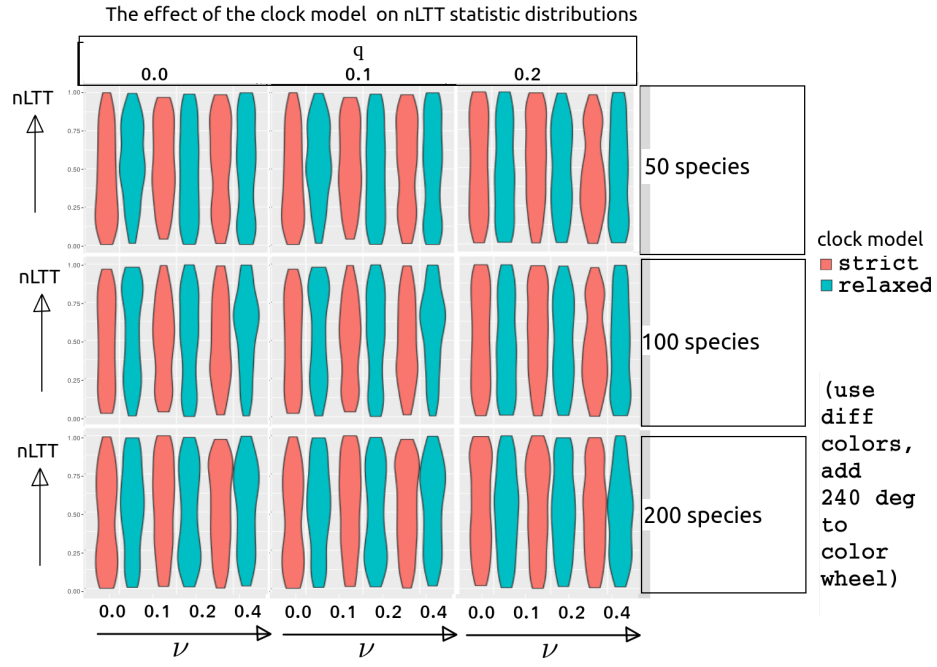


Figure 2: nLTT statistic distribution per biological parameter set per clock model, using the general data set, under the (correct) assumption of a Jukes-Cantor site model.

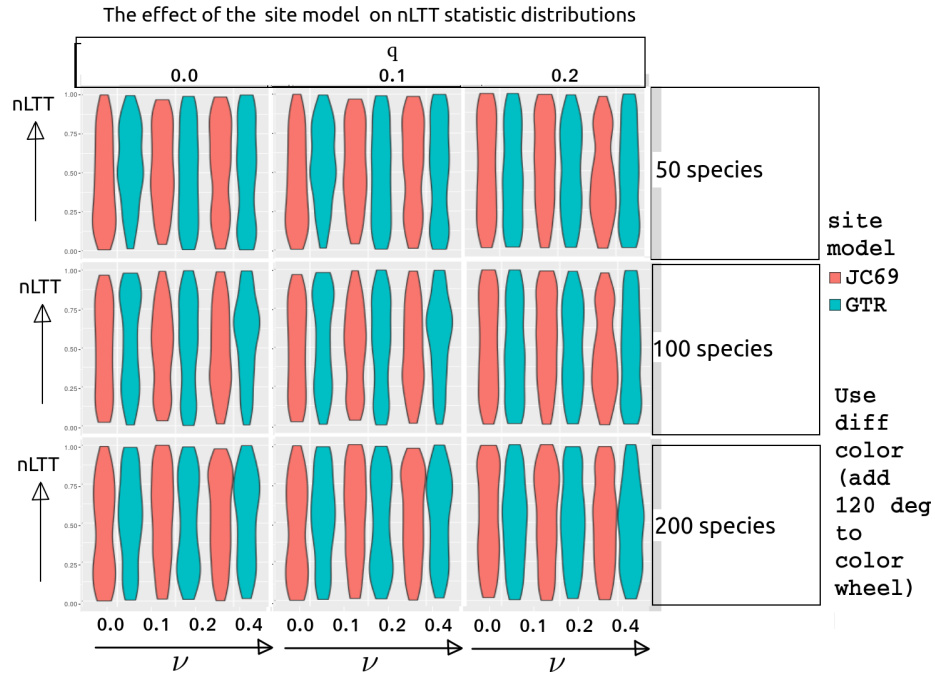


Figure 3: nLTT statistic distribution per biological parameter set per site model, using the general data set, under the (correct) assumption of a strict clock model.

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. GL created and tested the MBD

	Description	Values
λ	Per-species speciation rate	calculated
μ	Per-species extinction rate	0.0, 0.1
ν	Multiple speciation trigger rate	occurs never, once, twice, four and eight times
q	Per-species probability of multiple speciation	0, 0.25, 0.5, 1.0
n	Number of good taxa	50, 100, 200
t_c	Crown age	15
σ_c	Standard deviation around crown age	0.001
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 MBD tree generation	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 are the MBD model’s parameters. The RNG seed R_i is 1 for the first simulation, 2 for the next, and so on. 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).

n	Description
12 [RJCB: recalc]	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 **[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 4: Overview of the 11 parameters estimated by BEAST2

254 package. RJCB created and tested the experiment. GL and RJCB wrote the
255 first draft of the manuscript. RSE contributed substantially to revisions.