

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 26, 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 [GL: I've taken the freedom to change your proposed structure.
51 Feel free to change whatever you don't like it. I might be not com-
52 pletely accurate on the biological background.] [RJCB: It's just fine
53 :-)]

54 Current phylogenetic tools assume that only a single speciation event can
55 occur at the same time. While this assumption is useful to construct a wide
56 variety of successful models [RJCB: @gio: citation here], they disallow for
57 environmental changes that trigger speciation on a large scale, for example, the
58 cichlid fish diversification in the African Great Lakes: Malawi, Tanganyika and
59 Victoria [RJCB: @gio: citation here].

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

64 [GL: If I described the process in the same way you report in the
65 example I would probably end up writing the same things that we
66 say a few lines below, where we describe the parameters. Don't you
67 think?] [RJCB: You described the model in the Methods. I moved it

68 to here] In the MBD model, parameters λ and μ correspond, respectively, to
69 the usual per-species speciation and extinction rates. The model also introduces
70 two new parameters: ν is the total rate for an environmental change to trigger,
71 which leads to a potentially multiple speciation event [RJCB: How exactly?
72 When the environment change is triggered, is the speciation rate q or
73 $\lambda + q$? What ends the change?]. If such event triggers each species present
74 at that moment in time can undergo a speciation event with probability q .

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

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

96 The MBD model has 4 parameters, depicted in table 2.

97 [RJC: describe parameter values used here, example is below]
 98 [GL: @richel: I described the meaning of each parameter. I guess it
 99 is fine in this way. For the setting I think we have to wait to decide
 100 which values are actually worth a try] [RJC: I assume you have some
 101 idea of parameter values? For example, I think you will not use a λ

102 of -3 or 1234. Please put some reasonable values here, like I did in
 103 Figure 1 for you] We use [RJCB: please check me] $\nu = 0$ as our control
 104 for which the MBD model reduces to the BD model.

105 We simulate protracted birth-death trees, using the MBD package (Etienne
 106 2015) in the R programming language (R Core Team 2013). The first tree
 107 has a random number generator seed of 1, which is incremented by 1 for each
 108 simulated tree. For each combination of λ, μ, ν and q , we generate species trees
 109 with a crown age of 15 million years [GL: are we sure we wanna try 15
 110 million years. so far i've been trying only with 10 million years, which
 111 most of the time is working really well] [RJCB: I suggest yes. Most
 112 papers associated with Rampal use 15 million years, including my
 113 raket paper. Just use lower speciation rates]. Only trees with the desired
 114 number of good taxa are kept.

115 We create one data set to explore parameter space, All the trees with the
 116 correct number of good species are kept. Based on the species tree, we simu-
 117 late a DNA alignment that has the same history as this species tree, using the
 118 phangorn package (Schliep 2011). We set the nucleotides of the DNA alignment
 119 to follow a Jukes-Cantor (Jukes *et al.* 1969) nucleotide substitution model, in
 120 which all nucleotide-to-nucleotide transitions are equally likely. The DNA se-
 121 quence of the root ancestor consists of four equally sized single-nucleotide blocks
 122 of adenine, cytosine, guanine and thymine respectively. For example, for a DNA
 123 sequence length of 12, this would be AAACCCGGGTTT. The order of nucle-
 124 tides does not matter in this study, because we do not consider several partitions
 125 of the sequence with their own parameters. Only the frequency of occurrence
 126 matters. In our Bayesian inference (see below) we use the same site model as the
 127 (obviously correct) site model prior, but we also explore the effect of assuming a
 128 more complex site model prior. We predict with the more complex substitution

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

143 **[RJC: must rewrite, use pirouette as a starting point]** From an
 144 alignment, we run a Bayesian analysis and create a posterior distribution of
 145 trees and parameters using the `pirouette` (Bilderbeek 2018) package that sets
 146 the input parameters similar to BEAUti 2 and then runs BEAST2. For our site
 147 model, we assume either a Jukes-Cantor or GTR nucleotide substitution model.
 148 The Jukes-Cantor model is the correct one, as it is used for simulating that
 149 alignment, where the GTR model is the site model that is picked as a default
 150 by most users. For our clock model, we assume either a strict or relaxed log-
 151 normal clock model. Also here, the strict clock model is the correct one, as it is
 152 used for simulating the alignment, but the relaxed log-normal clock model is the
 153 one most commonly used. We set the BD model as a tree prior, as gauging the
 154 effect of this incorrect assumption is the goal of this study. We assume an MRCA
 155 prior with a tight normal distribution around the crown age, by choosing the

156 crown age as mean, and a standard deviation of $0.5 \cdot 10^{-3}$ time units, resulting
 157 in 95% of the crown ages inferred have the same resolution (of 10^{-3} time units)
 158 as the alignment. We ran the MCMC chain to generate 1111 states, of which we
 159 remove the first 10% (also called the 'burn-in'). Of the remaining 1000 MCMC
 160 states, the effective sample size (ESS) of the posterior must at least be 200
 161 for a strong enough inference (Drummond & Bouckaert 2015). An ESS can be
 162 increased by increasing the number of samples or decreasing the autocorrelation
 163 between samples. If the ESS is less than 200, we decrease autocorrelation by
 164 doubling the MCMC sampling interval of that simulation, until the ESS exceeds
 165 200.

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

173 The input trees generated with a [RJC: @gio: parameter that is set
 174 to reduce the MBD model to BD] allow us to measure the noise of the
 175 experiment.

176 We produce one data set as a comma-separated file. The general data set
 177 has 144 [RJC: recalc] different combinations of biological parameter com-
 178 binations, site and clock models. The data set to investigate sampling has 552
 179 [RJC: recalc] different combinations of biological parameter combinations,
 180 site models, clock models and sampling methods. The experiment is compu-
 181 tationally intensive: pilot experiments show that the experiment takes roughly
 182 100 days of CPU time and 20 days of wall clock time (which includes the queued

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

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.

For both data sets, we display the nLTT statistics distribution per biological parameter combination as a violin plot. We only show the nLTT distributions that were generated under the (correct) assumptions of a Jukes-Cantor site model and a strict clock model, separated per sampling method used. We display the nLTT statistic distributions separated per site or clock model in the supplementary information.

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. (2018) *pirouette: create a posterior from a phylogeny*.

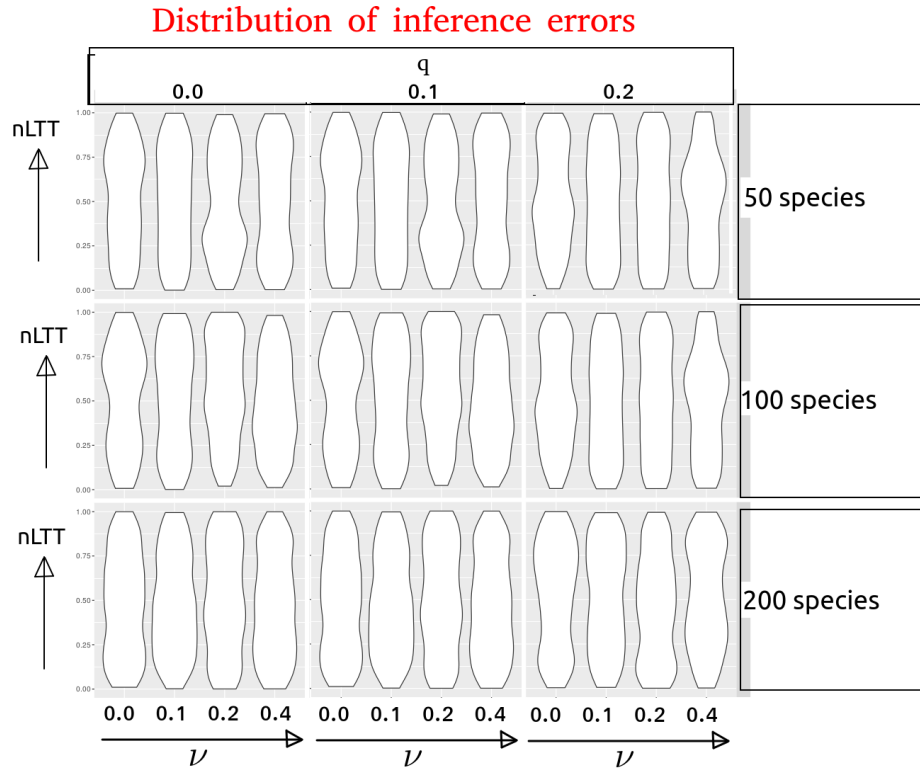


Figure 1: nLTT statistic distribution per biological parameter set, using the general data set, under the (correct) assumptions of a strict clock and Jukes-Cantor site model.

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

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

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

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

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

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

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

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

217 Janzen, T., Höhna, S. & Etienne, R.S. (2015) Approximate bayesian compu-
218 tation of diversification rates from molecular phylogenies: introducing a new
219 efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, **6**,
220 566–575.

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

- 223 Möller, S., du Plessis, L. & Stadler, T. (2018) Impact of the tree prior on
 224 estimating clock rates during epidemic outbreaks. *Proceedings of the National*
 225 *Academy of Sciences*, p. 201713314.
- 226 Posada, D. & Buckley, T.R. (2004) Model selection and model averaging in
 227 phylogenetics: advantages of akaike information criterion and bayesian ap-
 228 proaches over likelihood ratio tests. *Systematic biology*, **53**, 793–808.
- 229 R Core Team (2013) *R: A Language and Environment for Statistical Computing*.
 230 R Foundation for Statistical Computing, Vienna, Austria.
- 231 Rannala, B. & Yang, Z. (1996) Probability distribution of molecular evolution-
 232 ary trees: a new method of phylogenetic inference. *Journal of molecular*
 233 *evolution*, **43**, 304–311.
- 234 Schliep, K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**,
 235 592–593.
- 236 Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of
 237 dna sequences. *Lectures on mathematics in the life sciences*, **17**, 57–86.
- 238 Yang, Z. & Rannala, B. (2005) Branch-length prior influences bayesian posterior
 239 probability of phylogeny. *Systematic Biology*, **54**, 455–470.

240 A Acknowledgements

241 [RJC: put this section here, as the journal does not request for this]

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

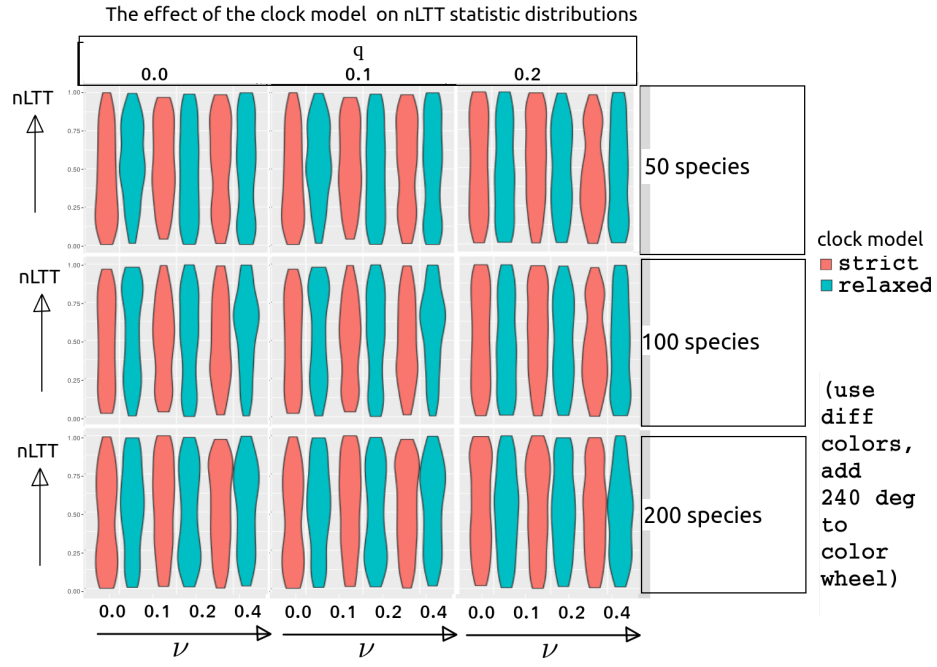


Figure 2: nLTT statistic distribution per biological parameter set per clock model, using the general data set, under the (correct) assumption of a Jukes-Cantor site model.

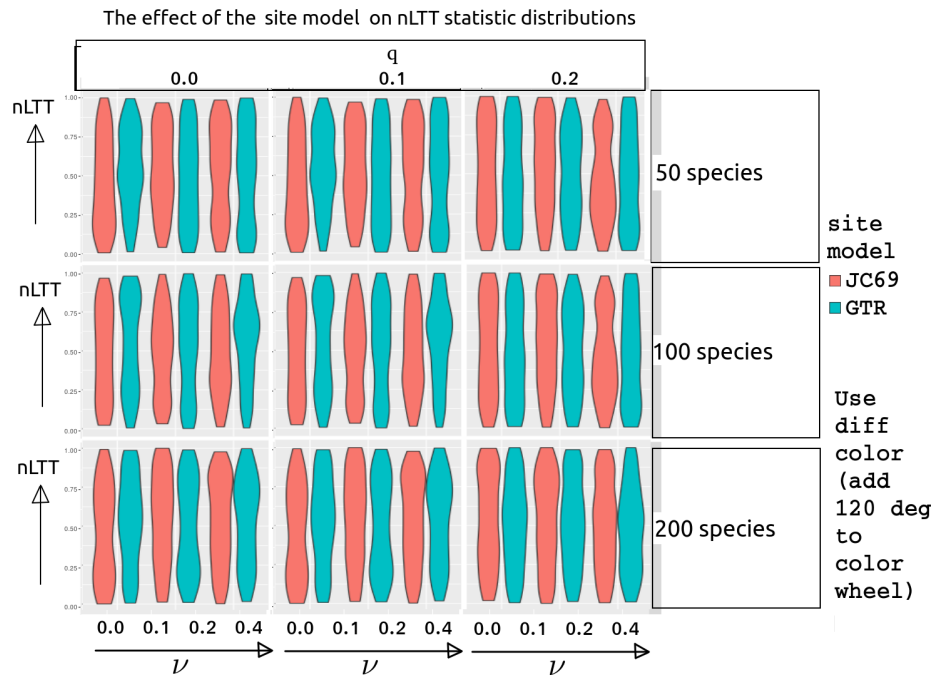


Figure 3: nLTT statistic distribution per biological parameter set per site model, using the general data set, under the (correct) assumption of a strict clock model.

	Description	Values
λ	[RJCB: @gio: MBD params here]	0.1, 0.3, 1.0, 10^9
λ	Speciation rate	0.1, 0.3, 1.0, 10^9
μ	Extinction rate	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. 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

245 B Authors' contributions

246 [RJC*B*: put this section here, as the journal does not request for this]
247 RSE [RJC*B*: @gio: I assume this this is true?] conceived the idea for
248 this experiment. GL created and tested the MBD package. RJC*B* created and
249 tested the experiment. GL and RJC*B* wrote the first draft of the manuscript.
250 RSE contributed substantially to revisions.