



LJMU Research Online

Brown, RP and Jin, Y

Partition Number, Rate Priors and Unreliable Divergence Times in Bayesian phylogenetic dating

<http://researchonline.ljmu.ac.uk/id/eprint/7328/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Brown, RP and Jin, Y (2017) Partition Number, Rate Priors and Unreliable Divergence Times in Bayesian phylogenetic dating. Cladistics. ISSN 0748-3007

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

<http://researchonline.ljmu.ac.uk/>

1

2

Title Page

3 Partition Number, Rate Priors and Unreliable Divergence Times in Bayesian phylogenetic
4 dating

5 Yuanting Jin¹, Richard P. Brown^{2*}

6 ¹*College of Life Sciences, China Jiliang University, Hangzhou, 310018, P. R. China, e-mail*
7 *jinyuanting@126.com*

8 ²*School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool L3,*
9 *3AF, UK.*

10 **Correspondence*

11 *Running title.- RATE PRIORS AND BEAST DIVERGENCE TIME DATING*

12

13

Abstract

14 More loci/partitions should improve Bayesian estimation of divergence times on
15 phylogenies but it has recently been shown that this can lead to surprisingly poor estimation
16 due to the way it affects the prior on mean substitution rate. Here we consider the likely
17 impact of partition number on divergence times analyses carried out using the program
18 BEAST. Mitochondrial genome data from toad-headed lizards (genus *Phrynocephalus*) from
19 the Qinghai-Tibetan Plateau were used to examine this effect. Under increased partitioning
20 of the sequences, BEAST posterior divergence times became unreasonably narrow and
21 downwardly biased due to misspecification of the mean substitution rate prior. This effect
22 was detectable when relatively few partitions were used (i.e., between four and eight), but
23 became very acute for 27-86 partitions. Fortunately, a correction that adjusts the standard
24 deviation of the mean of locus rates led to results that were equivalent to those obtained
25 using the latest version of the program MCMCtree, which implements a new
26 gamma-Dirichlet prior to overcome this problem. A review of the literature shows that a
27 substantial number of BEAST dating studies are likely to have been affected by this
28 misspecification of the rate prior.

29

30

Table of contents

31 Introduction

32 Analyses of Mitgenome Sequences

33 Dependence of Divergence Times on Number of Loci

34 Acknowledgments

35 References

36

37 **Introduction**

38 Bayesian estimation of divergence times on a phylogeny has been the subject of intensive
39 research for over ten years, see Yang (2014) and Drummond and Bouckaert (2015).
40 Nonetheless, many statistical aspects of Bayesian dating are still under investigation with
41 the impact of tree and rates priors on posterior times being a particularly active area (Brown
42 & Yang, 2010; Dos Reis *et al.*, 2014; Heled & Drummond, 2012; Ritchie *et al.*, 2017).

43 Theoretical work has clarified the relationships between decreased posterior interval
44 widths on divergence times and both increased amounts of sequence data and number of
45 loci/partitions, as well as demonstrating how this improvement is limited by uncertainties in
46 the calibrations (Rannala & Yang, 2007; Zhu *et al.*, 2015). Note that here we use the terms
47 partition and locus synonymously to define sequence alignments to which individual models
48 are applied. Some partitioning effects have been explored using maximum likelihood and
49 Bayesian dating of large amounts of nuclear sequence (Mulcahy *et al.*, 2012). More recently
50 a significant effect was identified where increasing the number of data partitions led to
51 misspecification of the prior on locus rates (described below). This problem was addressed
52 by incorporating new priors in MCMCtree (v4.8)(Dos Reis *et al.*, 2014), a program which
53 dates sequence divergence on a fixed topology. Using newly-generated mitogenome data
54 from Chinese *Phrynocephalus* lizards, we examine the potential impact of the rate prior
55 misspecification in studies that have used a very widely-used alternative program: BEAST.

56 Bayesian dating analyses generally treat locus rates as independent and identically
57 distributed (i.i.d) random variables which are typically specified from gamma or lognormal

58 distributions. Individual rate priors can strongly influence divergence time estimation
59 because the mean of the locus rates under a strict clock (or the mean of the mean branch
60 rate under an independent-rates relaxed clock) will have a decreasing standard deviation as
61 more locus rates are sampled. The standard deviation of the mean of the rate across loci is
62 $s/\sqrt{n_L}$ (where s is the standard deviation of the locus rates and n_L is the number of loci) and
63 so tends to zero as the number of loci tends to infinity (Dos Reis *et al.*, 2014). Hence the
64 mean locus rate prior becomes very restrictive which, due to the confounding of rate and
65 time, leads to overly-narrow and biased posteriors on divergence times. In other words, as
66 the number of loci/partitions increase, posteriors will provide the misleading impression
67 that divergence times are known with a high degree of precision and the location of the
68 posteriors will be inaccurate because they will be heavily influenced by the restrictive mean
69 locus rates prior.

70 New gamma-Dirichlet priors on locus rates and variance of log-transformed rates, σ^2 ,
71 have been implemented in MCMCtree (v4.8) to overcome the misspecification of the mean
72 locus rate prior (Dos Reis *et al.*, 2014). An alternative option that has been proposed for
73 other programs is to proportionally increase the variances of the individual rate priors in a
74 way that holds constant the standard deviation of the mean locus rate prior. Dos Reis *et al.*
75 (2014) suggested modification of the shape (α) and scale (β) parameters of the gamma prior
76 on locus rates, to $G(\alpha/n_L, n_L/\beta)$, where n_L is the number of loci. The variance of the mean
77 locus rates prior will then be the same as for a one partition analysis with the locus rate
78 specified from $G(\alpha, \beta)$.

79 Many BEAST divergence time analyses have been published in the past five years alone,
80 and partitioning of data from one marker and/or using multiple loci appears common. Here
81 we consider the likely impact of misspecification of the prior on mean rate on these
82 divergence time estimates. This is assessed using sequence data obtained by ourselves and
83 others from the mitochondrial genomes of Chinese *Phrynocephalus* lizards from the
84 Qinghai-Tibetan Plateau (QTP). We then consider the effects of the correction proposed by
85 (Dos Reis *et al.*, 2014) on BEAST analyses and compare it with the new gamma-Dirichlet
86 prior in MCMCtree.

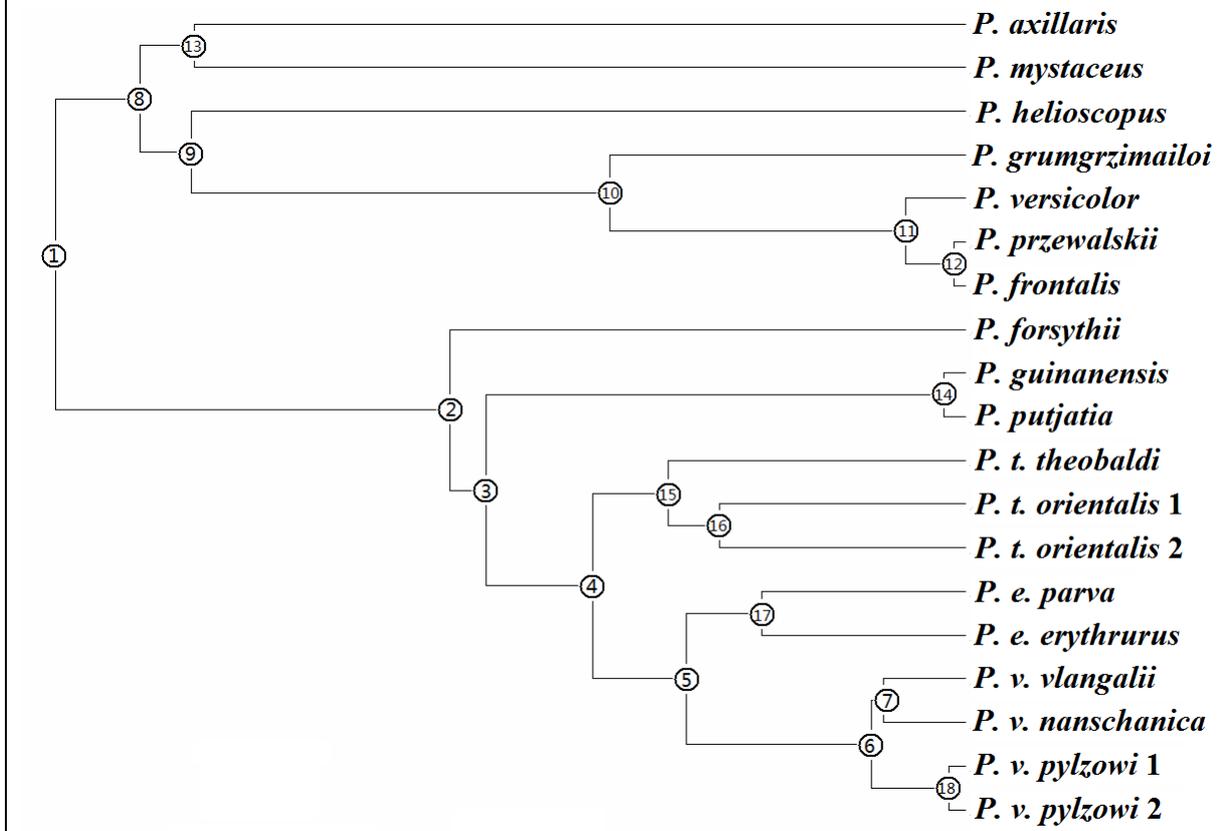
87 **Analyses of Mitgenome Sequences**

88 We analyzed 19 mitochondrial genomes from 13 recognized QTP *Phrynocephalus* with
89 intraspecific sampling of three of these: *P. theobaldi* (3 subspecies/lineages), *P. vlangalii* (4
90 subspecies/lineages), *P. erythrurus* (2 subspecies). The full list of specimens and their
91 capture sites are listed in Supplementary file 1. The species form a monophyletic group and
92 are subdivided into reciprocally monophyletic viviparous groups and oviparous groups (Jin &
93 Brown, 2013). Of these, eight new mitochondrial genomes have been recently sequenced
94 and 11 published genomes were already available (all genomes are available on GENBANK:
95 see Supplementary file 1). The *Phrynocephalus* mitochondrial genome sequencing approach
96 is described in Liao and Jin (2016).

97

98 Figure 1. *Phrynocephalus* tree topology.

99 The *Phrynocephalus* tree topology used in all analyses, with node labels.



100

101 Our dating analyses used a single topology derived from the tree previously inferred
102 from mtDNA and nuclear sequences (Jin & Brown, 2013; Fig. 1). Bayesian and NJ analyses of
103 the current mitogenome data did not reveal any discordance with this tree. The
104 mitogenome sequence alignment was divided using 11 different strategies that provided
105 between 1 and 86 sequence partitions. Although many studies now tend to use automated
106 methods of finding partitions, the main focus here was the impact of the number of
107 partitions rather than the partition characteristics. In brief, the strategies were based on
108 both mitochondrial genes and different positions within these genes: codon position in
109 protein-coding genes, stem or loop regions of rRNAs and tRNAs. Fewer partitions were

110 achieved by grouping genes and/or grouping codon positions and/or grouping stem/loop
111 RNA regions. For example, the eight partition analysis grouped all protein coding genes but
112 divided the sequences by codon positions, tRNA regions were grouped and sequences
113 divided by stem/loop, rRNA regions were treated as for the tRNAs while the final partition
114 was the control region. Analyses with higher numbers of partitions made use of all possible
115 divisions. For example, the 86 partition analysis divided the three codon positions for each of
116 the 13 genes, stems or loops for each of the 22 tRNA and two rRNA genes and the control region
117 sequence (note that two very short tRNA partitions of 25 and 27 bp with negligible
118 information content were excluded from the 75, 84 and 86 partition analyses). The
119 partitions are described fully in Supplementary file 2.

120 All data partition strategies were analyzed using both BEAST (v.1.8.1) and MCMCtree
121 (v.4.8). To ensure the generality of our findings we also repeated the one, four, 27 and 46
122 partition analyses using BEAST 2 (v. 2.4.7)(Bouckaert et al., 2014) with replicate
123 specifications to those described below for the BEAST 1.8.1 uncorrected gamma rate prior
124 analyses.

125 Preliminary analyses revealed that the most generally suitable site model was HKY+G,
126 which is available in both programs, and therefore applied independently to each partition.
127 A relaxed clock with uncorrelated rates on branches drawn from a lognormal distribution
128 was also applied independently to each partition.

129 The same node age calibrations were used in both programs and have been justified
130 previously (Jin & Brown, 2013). The age (Ma) of the node that was ancestral to all *P*.

131 *vlangalii* and *P. erythrurus* (node 5, Fig. 1) was specified from the uniform distribution
132 U(1.35-5.00) and the node that was ancestral to all of the oviparous species (node 8, Fig. 1)
133 was specified from U(7.24-10.95). A maximal constraint of 25 Ma was placed on the root.
134 One of the differences between programs is that the upper and lower limits of the uniform
135 distribution are hard in BEAST, but are soft in MCMCtree. The latter implements an
136 exponential decline in density above and below the specified limits of the distribution (here,
137 each tail comprised 2.5% of the total density).

138 The prior on rates in MCMCtree was a flexible gamma prior in which both shape (α) and
139 scale (β) parameters were 1, denoted as G(1,1). The gamma distributions are specified in
140 MCMCtree using shape/rate rather than shape/scale parameterization but to be consistent
141 we describe all gamma distributions in terms of the latter. The G(1,1) distribution provides a
142 flexible prior for substitution rates (95% Highest posterior density (HPD): 0.025-3.689
143 subs/site/Ma) and was also used for the σ^2 prior on rate variation and the α shape prior. A
144 G(5,1) prior was specified for κ , the transition: transversion rate ratio. The
145 Birth-Death-Sampling prior on times was used with parameters $\lambda=5$, $\mu=5$, $\rho=0.1$, as this has
146 been shown to be quite flexible (Brown & Yang, 2010).

147 All BEAST analyses were all carried out on the fixed topology (Fig. 1) to replicate
148 MCMCtree analyses. A first set of “uncorrected” BEAST analyses specified locus rates
149 through the ucl.d.Mean parameters from a G(1,1) distribution (for all partitions). A second
150 set of “corrected” analyses applied variance corrections to this gamma prior for analyses
151 with ≥ 2 partitions as proposed by Dos Reis *et al.* (2014): priors were specified from G(α/n_L ,

152 n_L/β), which simplifies to $G(1/n_L, n_L)$ here. Dos Reis et al. (2014) also implemented a new
153 prior on the variance of the log transformed rates, which could be emulated through
154 corrections to `uclid.Stdev` parameters in BEAST, but we did not attempt this. The prior on
155 times was sampled from a Birth-Death speciation prior which has two parameters:
156 speciation rate, specified from the uniform distribution $U(0,10000)$, and relative death rate,
157 specified from $U(0,1)$. (An example BEAST input file is provided in Supplementary file 3).
158 Prior distributions were estimated by repeating analyses without data.

159 **Dependence of Divergence Times on Number of Loci**

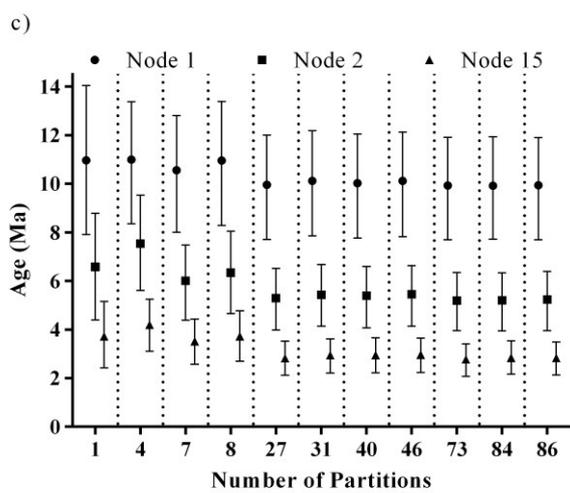
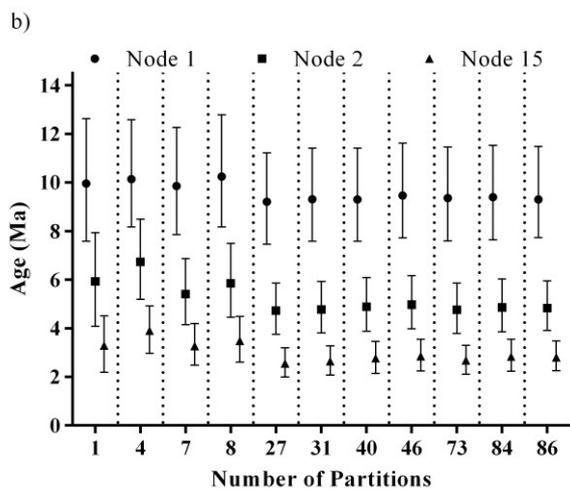
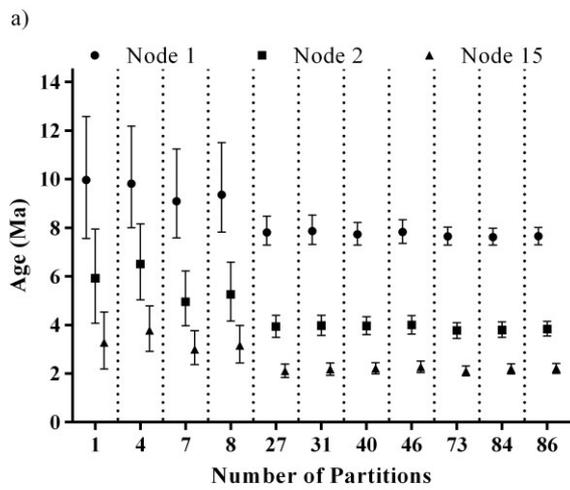
160 Uncorrected BEAST analyses suffered from the general and major problem described by Dos
161 Reis *et al.* (2014) for MCMCtree. Both the locations and widths of posterior divergence
162 times were highly dependent on the number of partitions (Fig. 2a). Increasing numbers of
163 partitions led to unreasonably narrow posteriors with lower median divergence times. We
164 confirmed this effect is not confined to BEAST v. 1.8: the replicate BEAST 2.4.7 analyses gave
165 the same means and posterior widths to those obtained from the earlier version of the
166 program. Despite a relatively recent root (~ 10 Ma), posterior means at many nodes were
167 generally 1-2 Ma lower for analyses with more than 8 partitions compared to analyses with
168 no data partitioning. At the same time, the widths of the 95% Highest Posterior Densities
169 (HPD) showed drastic decreases, with the interval on the root decreasing from
170 approximately 5 Ma to 0.7 Ma. The change in the mean depends on the degree of
171 misspecification of the priors on rates but underestimation of the uncertainty in divergence

172 times is a general problem. The effects are noticeable even for quite low numbers of
173 partitions.

174

175 Figure 2. Posterior divergence times obtained from BEAST and MCMCtree.

176 Posterior divergence times (means and 95% HPDs) at three selected nodes (1, 2 and 15) on
177 the *Phrynocephalus* tree for different numbers of partitions. a) BEAST analyses with a G(1,1)
178 prior on all rates, b) BEAST analyses with corrected priors on rates, c) MCMCtree v4.8
179 analyses. BEAST prior divergence times (95% HPDs in Ma) were (7.3-18.1) for node 1,
180 (3.6-13.8) for node 2, and (0.2-5.7) for node 15. These priors on times are not affected by
181 specification of the i.i.d priors on rates or the number of partitions.



184 The impact of increasing numbers of partitions was greatly ameliorated in BEAST by
185 proportionately increasing the variances of the i.i.d. priors on individual partition rates
186 relative to the number of partitions (Fig. 2b). This correction had no effect on priors on
187 divergence times but maintained the standard deviation of the mean locus rate prior
188 constant for analyses with different numbers of partitions. The success of the correction
189 was verified by the similarity with results from MCMCtree (Fig. 2c), which incorporates a
190 new gamma-Dirichlet prior to overcome misspecification of the mean locus rate prior. There
191 were some differences in posterior divergence times between MCMCtree and the corrected
192 BEAST analyses, but this would be expected due to several significant differences between
193 the programs, including the way calibrations are specified.

194 As expected, posterior intervals on divergence times in correctly-specified MCMCtree
195 and BEAST analyses were narrower with increased numbers of loci: posterior variances in
196 divergence times are expected to decrease at the rate $1/n_L$ (Zhu et al., 2015). This is
197 exemplified by the width of the posterior on the root: 5 Ma with no partitioning compared
198 with 3.8 Ma for 86 partitions. Decreases in respective widths with increasing numbers of
199 partitions were relatively greater in some other nodes (Fig. 2). This finding underlines the
200 advantage of using more loci, providing the mean locus rate prior is correctly specified.

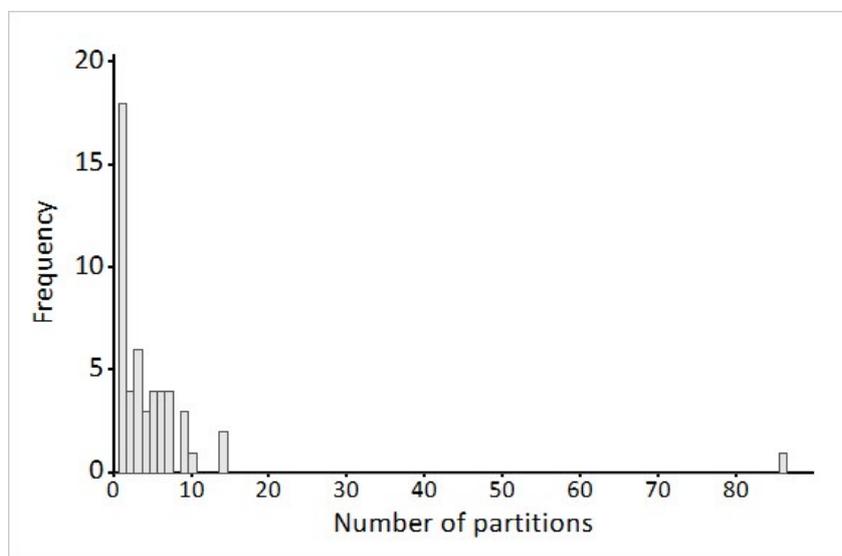
201 It is more difficult to explain the more subtle pattern of variation in the locations of
202 posterior ages in the corrected BEAST (and MCMCtree) analyses with increasing number of
203 loci. Mean ages were younger for fewer partition analyses. For example, the root was 0.7
204 Ma younger (and some other basal nodes were up to 2 Ma younger) when the data were

205 not partitioned, compared with 27-86 partitions (Fig. 2). Identification of the priors that
206 might be responsible for this change is not straightforward. The fact that the same
207 between-partition pattern is seen in MCMCtree shows that this effect is general, rather than
208 being specific to BEAST. An increasingly influential prior on times with an increase in
209 uninformative partitions seems an unlikely explanation (and in fact the pattern runs counter
210 to this). An alternative explanation is that it is due to the influence of individual locus rate
211 priors on relatively uninformative partitions. The gamma distributions are flexible but have a
212 mean ($\mu=1$) that must exceed the partition rates and so some/all posterior branch rates will
213 be overestimated when phylogenetic information is lacking. In relatively uninformative
214 partitions the gamma prior will be very influential and the overestimated ucl. Mean rate
215 will lower divergence times due to the confounding of time and rate. This explanation
216 provides a better fit to the pattern observed in our analyses and was supported by: i)
217 simulation and analyses of datasets that contained non-informative and informative
218 partitions, and ii) generally higher posterior means of mean locus rates in MCMCtree for
219 greater numbers of partitions.

220 The following search terms were used in the search engine Bing/Academic: "beast",
221 "divergence time", and "dating" to find relevant papers published between 2007-2017. The
222 search produced 15500 hits which were ordered in terms of their suitability to the search
223 term. We sampled the first 50 papers/theses that appeared to represent independent
224 BEAST divergent time analyses of empirical data and recorded the number of partitions
225 used. The numbers of partitions in these studies ranged from 1-86 (mean = 5.6, median
226 =3.0) summarized in Figure 3. A significant proportion of these studies (38%) used five or

227 more partitions. It is likely that some of these studies linked the clock across all partitions
228 (this information was frequently missing), in which case the problematical mean locus rate
229 prior should not affect divergence times in the way described here, as found by Zheng and
230 Wiens (2016). Nevertheless, we conclude that a significant number of published BEAST
231 divergence time estimations are likely to have been affected by the prior on rates.
232 Application of the correction to the ucl.Mean prior on rates will however remove this issue
233 from future BEAST studies. This problem will also be helped by the development of methods
234 to assess the number of clock models that are suitable for a dataset (Duchêne *et al.*, 2014)
235 as this will likely lead to a reduction in the number of “clock-partitions” that are used in an
236 analysis.

237 Figure 3. Numbers of partitions used in a sample of 50 divergence time studies.



238

239 Acknowledgments

240 This work was supported by the National Natural Science Foundation of China (31372183).

241 We thank China Jiliang University for a visiting scholar award to RPB. Several members of

242 YTs research group contributed to laboratory or fieldwork including Xiaolong Tang (sample
243 collection), Haojie Tong, Qian Wang and Liufang Zhu. We thank Mario dos Reis and an
244 anonymous reviewer for their comments on an earlier version of the manuscript.

245 **References**

- 246 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut,
247 A. and Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian
248 evolutionary analysis. *PLoS Computational Biology*, 10(4), p.e1003537.
- 249 Brown, R. P., & Yang, Z. (2010). Bayesian dating of shallow phylogenies with a relaxed clock.
250 *Systematic Biology*, 59: 119-131.
- 251 Dos Reis, M., Zhu, T., & Yang, Z. (2014). The impact of the rate prior on Bayesian estimation
252 of divergence times with multiple loci. *Systematic Biology*, 63, 555–565.
- 253 Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with Beast*:
254 Cambridge University Press.
- 255 Duchêne, S., Molak, M., & Ho, S. Y. W. (2014). Clockstar: Choosing the number of
256 relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics*, 30,
257 1017-1019.
- 258 Heled, J., & Drummond, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and
259 divergence time estimation. *Systematic Biology*, 61, 138-149.
- 260 Jin, Y., & Brown, R. P. (2013). Species history and divergence times of viviparous and
261 oviparous Chinese toad-headed sand lizards (*Phrynocephalus*) on the
262 Qinghai-Tibetan plateau. *Molecular Phylogenetics and Evolution*, 68, 259-268.
- 263 Liao, P., & Jin, Y. (2016). The complete mitochondrial genome of the toad-headed lizard
264 subspecies, *Phrynocephalus theobaldi orientalis* (Reptilia, Squamata, Agamidae).
265 *Mitochondrial DNA Part A*, 27, 559-560.
- 266 Mulcahy, D. G., Noonan, B. P., Moss, T., Townsend, T. M., Reeder, T. W., Sites, J. W., et al.
267 (2012). Estimating divergence dates and evaluating dating methods using
268 phylogenomic and mitochondrial data in squamate reptiles. *Molecular Phylogenetics
269 and Evolution*, 65, 974-991.
- 270 Rannala, B., & Yang, Z. (2007). Inferring speciation times under an episodic molecular clock.
271 *Systematic Biology*, 56, 453-466.
- 272 Ritchie, A. M., Lo, N., & Ho, S. Y. (2017). The impact of the tree prior on molecular dating of
273 data sets containing a mixture of inter-and intraspecies sampling. *Systematic
274 Biology*, 66, 413-425.
- 275 Yang, Z. (2014). *Molecular evolution: A statistical approach*: Oxford University Press.
- 276 Zheng, Y., & Wiens, J. J. (2016). Combining phylogenomic and supermatrix approaches, and
277 a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52
278 genes and 4162 species. *Molecular Phylogenetics and Evolution*, 94, 537-547.

279 Zhu, T., Dos Reis, M., & Yang, Z. (2015). Characterization of the uncertainty of divergence
280 time estimation under relaxed molecular clock models using multiple loci. *Systematic*
281 *Biology*, 64, 267-280.

282