1

2 **Title Page**

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

4 dating

5 Yuanting Jin[1], 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

Abstract

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

30 **Table of contents**

36

Introduction

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

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

Bayesian dating analyses generally treat locus rates as independent and identically 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 $G(\alpha, \beta)$.

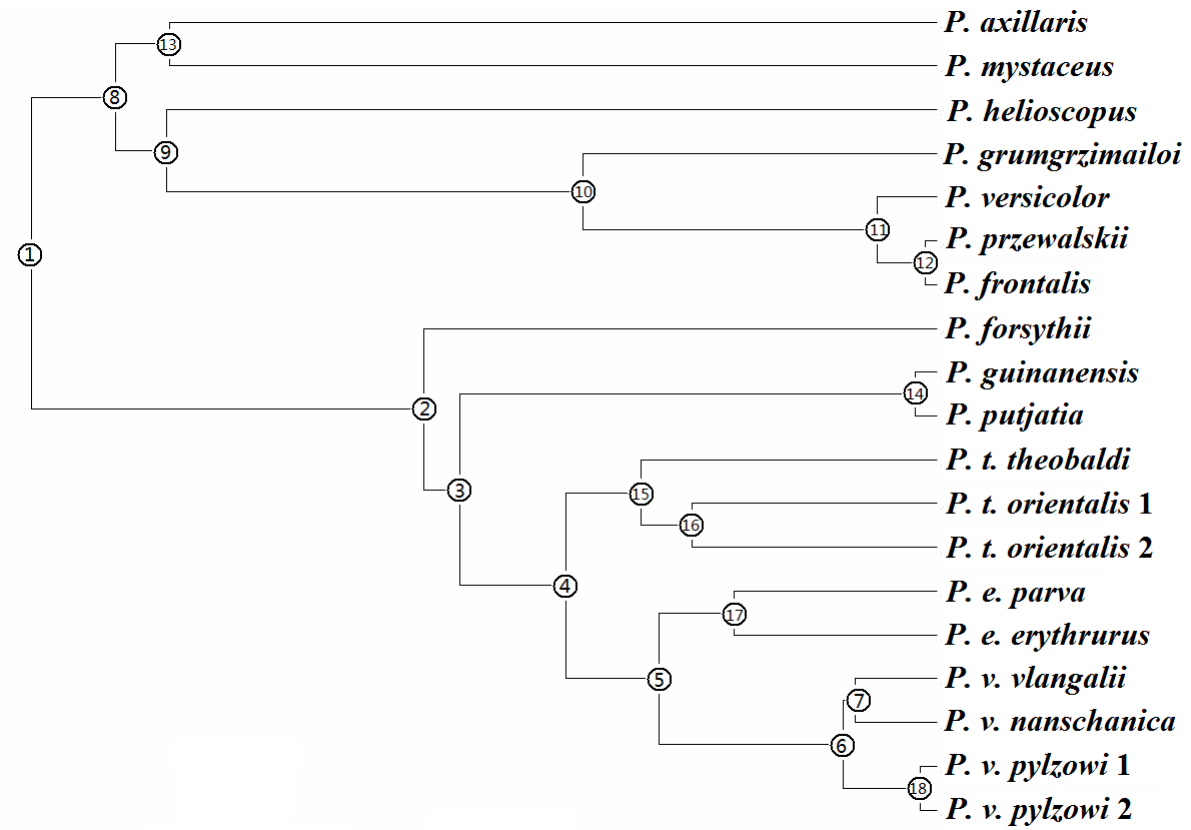
Many BEAST divergence time analyses have been published in the past five years alone, and partitioning of data from one marker and/or using multiple loci appears common. Here we consider the likely impact of misspecification of the prior on mean rate on these divergence time estimates. This is assessed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

Analyses of Mitgenome Sequences

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

Figure 1. *Phrynocephalus* tree topology.

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



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

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

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

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

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

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

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

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

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

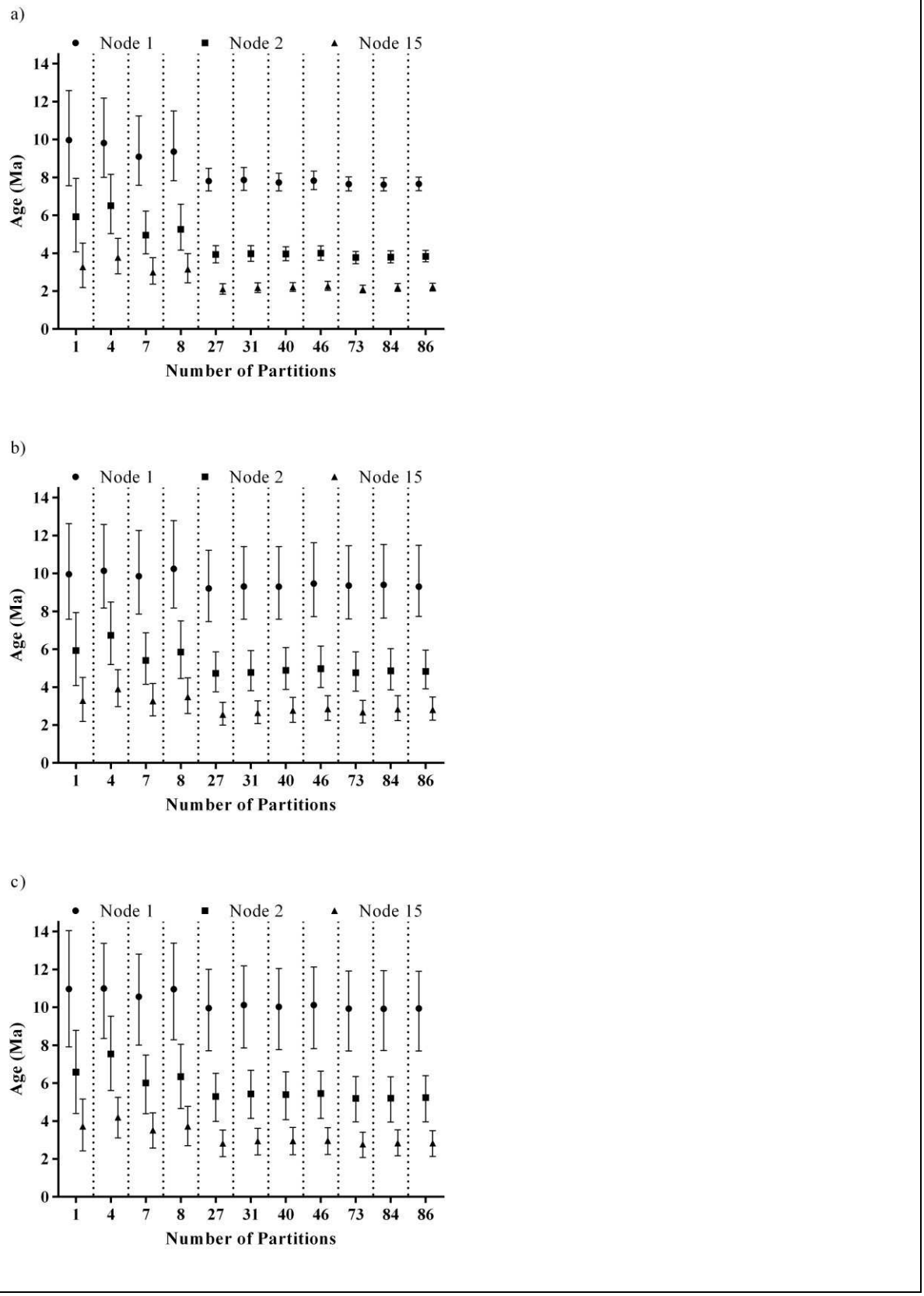
Dependence of Divergence Times on Number of Loci

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

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

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

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


a)
Node 1
Node 2
Node 15
Age (Ma)
0
2
4
6
8
10
12
14
1
4
7
8
27
31
40
46
73
84
86
Number of Partitions
b)
Node 1
Node 2
Node 15
Age (Ma)
Number of Partitions
c)
Node 1
Node 2
Node 15
Age (Ma)
Number of Partitions

182

183

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.

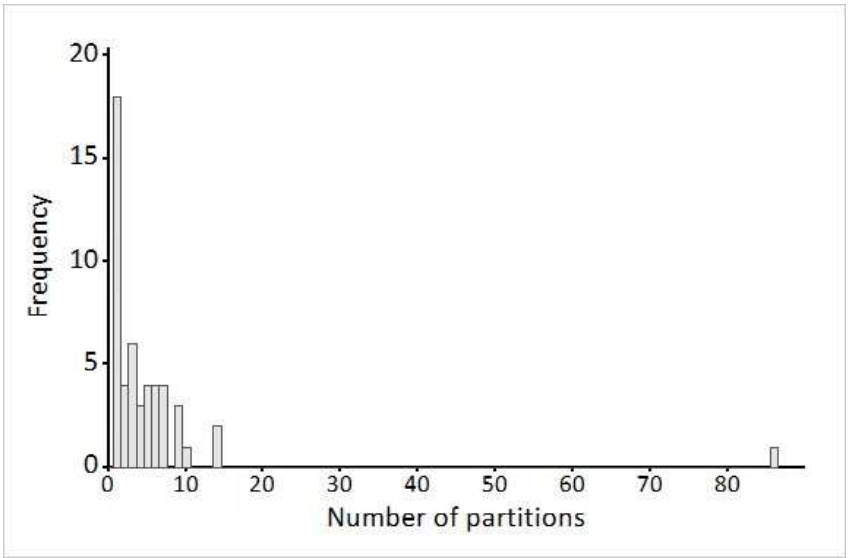
201 It is more difficult to explain the more subtle pattern of variation in the locations of
202 posterior ages in the corrected BEAST (and MCMCtree) analyses with increasing number of
203 loci. Mean ages were younger for fewer partition analyses. For example, the root was 0.7
204 Ma younger (and some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

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

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

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

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



Acknowledgments

This work was supported by the National Natural Science Foundation of China (31372183). We thank China Jiliang University for a visiting scholar award to RPB. Several members of

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

References

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A. and Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4), p.e1003537.

Brown, R. P., & Yang, Z. (2010). Bayesian dating of shallow phylogenies with a relaxed clock. *Systematic Biology*, 59: 119-131.

Dos Reis, M., Zhu, T., & Yang, Z. (2014). The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Systematic Biology*, 63, 555–565.

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with Beast*: Cambridge University Press.

Duchêne, S., Molak, M., & Ho, S. Y. W. (2014). Clockstar: Choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics*, 30, 1017-1019.

Heled, J., & Drummond, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology*, 61, 138-149.

Jin, Y., & Brown, R. P. (2013). Species history and divergence times of viviparous and oviparous Chinese toad-headed sand lizards (*Phrynocephalus*) on the Qinghai-Tibetan plateau. *Molecular Phylogenetics and Evolution*, 68, 259-268.

Liao, P., & Jin, Y. (2016). The complete mitochondrial genome of the toad-headed lizard subspecies, *Phrynocephalus theobaldi orientalis* (Reptilia, Squamata, Agamidae). *Mitochondrial DNA Part A*, 27, 559-560.

Mulcahy, D. G., Noonan, B. P., Moss, T., Townsend, T. M., Reeder, T. W., Sites, J. W., et al. (2012). Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. *Molecular Phylogenetics and Evolution*, 65, 974-991.

Rannala, B., & Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Systematic Biology*, 56, 453-466.

Ritchie, A. M., Lo, N., & Ho, S. Y. (2017). The impact of the tree prior on molecular dating of data sets containing a mixture of inter-and intraspecies sampling. *Systematic Biology*, 66, 413-425.

Yang, Z. (2014). *Molecular evolution: A statistical approach*: Oxford University Press.

Zheng, Y., & Wiens, J. J. (2016). Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 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