Molecular Clock Dating of Deep-Time Evolution Using Complex Mixture Models

Sishuo Wang[1,*], Andrew Meade[2]

1. Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China
2. School of Biological Sciences, University of Reading, Reading, UK

* Corresponding author: sishuowang@hotmail.ca

Abstract

Molecular clocks are a fundamental technique in evolutionary biology for establishing the timing and tempo of organismal divergence. However, currently available molecular clock methods, which often rely on simple homogeneous substitution models, can produce inaccurate time estimates, particularly for deep-time or rapidly evolving lineages where substitution heterogeneity and saturation are common. Hereby, we introduce phyloHessian (https://github.com/evolbeginner/phyloHessianWrapper), a Julia-based software to enable the use of complex mixture substitution models in molecular dating. phyloHessian employs different algorithms to improve the numerical accuracy in calculating the phylogenetic Hessian matrix, which is further integrated into PAML-MCMCtree's approximate likelihood framework to conduct dating analyses. Simulations mimicking phylogenies at different timescales demonstrate that complex mixture substitution models significantly enhance the accuracy of divergence time and substitution rate estimates in deep-time phylogenies. This pattern remains consistent across a wide range of uncertainties associated with molecular clock analysis. Additionally, mixture models display greater robustness to model and calibration specifications compared to their homogeneous counterparts. Empirical analysis of Microsporidia and Rickettsiales with different

substitution models shows that mixture models, compared to homogeneous models, yield accelerated substitution rates and in some cases significantly different divergence times, leading to a revised understanding of their host association origins. Our findings underscore the importance of incorporating complex mixture substitution models for constructing reliable evolutionary timelines and elucidating the evolutionary history of deep-time lineages.

Keywords

molecular clock, PAML, MCMCtree, complex mixture model, deep-time evolution, approximate likelihood

Introduction

Historically, our understanding of evolutionary time scales relied on limited fossil records, until a significant change occurred in the 1960s when molecular clock theory was proposed (Zuckerkandl and Pauling 1965), making it possible to infer evolutionary time and rate using genetic data. While the original concept of the molecular clock assumed that the rate of gene sequence changes is constant among lineages, further evidence shows that even the same gene can exhibit variations in rate among different species, leading to the development of relaxed molecular clock models which allow for lineage-specific evolutionary rate and a series of derivative models (Thorne et al. 1998; Drummond et al. 2006; Yang and Rannala 2006). Over years, molecular clock theory has evolved, continuously increasing in complexity and accuracy, providing a powerful tool for studying biological evolution. Particularly, with the development of Bayesian phylogenetic methods and the rapid growth of genetic sequence data, Bayesian models based on Markov chain Monte Carlo (MCMC) methods have become the most widely used molecular clock dating method, which allows for a comprehensive examination of various uncertainties in analysis, including factors such as calibrations and across-lineage rate differences (Ho and Duchêne 2014; dos Reis et al. 2016).

While molecular clock approaches offer a powerful framework for inferring evolutionary timescales, the reliability of such inferences heavily depends on accurately modeling the evolution of the molecular data, such as protein sequences, alongside the incorporation of calibration information (Guindon 2020). Protein evolution is typically modelled as site-homogeneous process using empirically derived substitution matrices of amino acid exchangeabilities (e.g., LG or WAG) and fixed equilibrium frequencies, with variation in evolutionary rates modeled with a gamma distribution (Yang 1993). However, while

computationally efficient, such models oversimplify real-world biological processes by failing to capture the heterogeneity in amino acid preferences driven by diverse structural and functional constraints across sites. To address these issues, site-heterogeneous substitution models have been proposed since (Lartillot and Philippe 2004; Pagel and Meade 2004). These models accounts for among-site substitution heterogeneity by incorporating predefined substitution matrices for different secondary structures and solvent accessibility (Le et al. 2008; Le and Gascuel 2010), across-site rate variation (Le et al. 2012), or mixtures of amino acid frequency profiles (Quang et al. 2008; Wang et al. 2008). Of particular interest is the last category, which captures different amino acid preferences across sites, as represented by the CAT model (Lartillot and Philippe 2004) and its simplified counterparts for maximum-likelihood phylogenetics called Cxx (where xx indicates 10 to 60 profile categories) (Quang et al. 2008). By better approximating the true biological complexity of amino acid substitution, profile mixture models mitigate artifacts like long branch attraction, thereby enhancing the ability to distinguish phylogenetic signal from homoplasy and displaying higher sensitivity to substitution maturation (Wang et al. 2018, 2019; Schrempf et al. 2020; Baños et al. 2024a). These features make profile mixture models especially well-suited for studying microbial evolution, where deep divergences and rapidly evolving sites are prevalent. Indeed, empirical analyses of deep-time phylogenomic data show that profile mixture models often exhibit superior model fit compared to homogeneous alternatives (Martijn et al. 2018; Williams et al. 2020; Galindo et al. 2021; Moody et al. 2022; Eme et al. 2023).

A particular outcome of using site-homogeneous models like LG+G is that branch lengths are usually underestimated in case of extensive sequence divergence (Venditti et al. 2008; Moody et al. 2022). This happens because proteins may not use all 20 amino acids at each position and

homogenous models incorrectly interpret this restricted use of amino acids as evidence of less evolutionary change than what truly occurred over long time (Lartillot et al. 2007; Wang et al. 2008). To see this, consider a site restricted to Asp and Glu. A mixture model would identify a preference for negatively charged amino acids and infer unobserved substitutions (e.g., Asp → Glu → Asp) even if two sequences show the same amino acid, leading to longer branch lengths. Obviously, this pattern becomes more readily observed with an increasing evolutionary scale. Studies have shown that the deepest branches of the tree of life, spanning approximately 4.0 billion years (Gyr), can display branch lengths up to twice as long under profile mixture models compared to the LG+G model (Moody et al. 2022; Wang and Luo 2025). Given that in molecular clock analysis branch length, the expected substitutions per site, is the product of time and (average) substitution rate of that branch (dos Reis et al. 2016; Guindon 2020), an important question emerges: do profile mixture models, which often excel in deep-time phylogenetic inference, lead to more accurate divergence time and substitution rate estimates?

To our knowledge, none of the available molecular clock programs have directly implemented complex substitution models including those profile mixture models, with PhyloBayes' CAT model being the exception—yet its high computational demands limit its use in large-scale molecular clock analyses. We recently developed bs_inBV (Wang and Luo 2025), enabling the use of diverse substitution models in the popular MCMCtree molecular dating software by applying a bootstrap approximation of the Hessian matrix, which captures likelihood surface curvature and is required by MCMCtree's fast approximate likelihood dating procedure (Thorne et al. 1998; dos Reis and Yang 2011). However, the bootstrapping process is computationally intensive, and its approximation sometimes inadequately represents the true Hessian. Additionally, and perhaps more importantly, the relative performance of different

substitution models in molecular clock dating remains unexamined through systematic simulation studies, leaving a critical gap in our understanding of how choice in substitution model affects divergence time estimation, particularly in deep-time or fast-evolving phylogenies.

Hereby, we introduce phyloHessian, a Julia-based software that uses numerical methods to calculate accurate Hessian matrices for diverse complex substitution models and integrates them directly into MCMCtree's approximate likelihood framework for rapid molecular clock dating. We perform extensive analyses on simulation data to comprehensively compare the performance of different substitution models in divergence time and substitution rate estimation under a wide range of conditions that mimic phylogenies from relatively recent to very ancient divergences. We further analyze empirical data to assess how different substitution models affect our understanding of ancient symbiont evolution and the origins of their host associations.

Results

phyloHessian: enabling molecular clock dating in MCMCtree with complex substitution models

To address the challenges in molecular clock analysis under various substitution models, we developed a Julia-based software phyloHessian. phyloHessian first calls IQ-Tree (Wong et al. 2025) or PhyML (Guindon et al. 2010), two widely used phylogenetic reconstruction software that implement a range of mixture models, to calculate the maximum-likelihood estimates (MLE) of branch length $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)^{\mathrm{T}}$ under user-specified substitution model. Next, it applies finite difference numerical methods to calculate the gradient $g = \left(\frac{\partial \ell(\hat{\theta})}{\partial \theta_1}, \dots, \frac{\partial \ell(\hat{\theta})}{\partial \theta_n}\right)^{\mathrm{T}}$ and Hessian matrix $H(\hat{\theta}) = \left(\frac{\partial^2 \ell(\hat{\theta})}{\partial \theta_i \partial \theta_j}\right)$, i.e., the first- and second-order derivatives of the log-

likelihood of the phylogenetic data evaluated at the MLEs. Last, the resulting Hessian matrix is integrated into MCMCtree to be used with its approximate likelihood method for divergence time and rate estimation. MCMCTree's approximate likelihood method approximates the log-likelihood surface for branch lengths using a Taylor expansion to the second order

$$\ell(\theta) \approx \ell(\hat{\theta}) + \mathrm{g}^{\mathrm{T}}\Delta\theta + \frac{1}{2}\Delta\theta^{\mathrm{T}}H(\hat{\theta})\Delta\theta,$$

where $\ell(\theta)$ denotes the log-likelihood as a function of branch lengths $\theta = (\theta_i = t_i r_i)$ expressed as a product of time spanning the two adjacent nodes and the absolute substitution rate.

To demonstrate the reliability of phyloHessian, we compared its results with those obtained from well-established phylogenetics software packages. First, we found that phyloHessian and IQ-Tree produced identical log-likelihoods under all tested complex substitution models available in IQ-Tree (Data S1). Further, we validated phyloHessian's Hessian calculation by confirming that the Hessian matrix obtained by phyloHessian and CODEML under site-homogeneous models (such as LG+G) were nearly identical (Data S2).

Profile mixture substitution models enhance time and rate estimation in deep-time and fast-evolving phylogenies: evidence from simulations

To evaluate the impact of mixture substitution models (Methods: Mixture substitution model) on divergence time and substitution rate estimates, and to compare their performance to profile homogeneous models like LG+G, we conducted extensive simulations. To simulate a range of evolutionary scenarios spanning relative recent to very deep-time divergences, we followed our prior study (Wang and Luo 2025) to generate 30 replicates of 20-tip birth-death trees with 300-amino acid (aa) sequences simulated under the substitution model LG+C60+G4{1.0} (hereafter LG+C60+G where "+G4{1.0}" specifies four rate category discrete Gamma distribution with

shape parameter α=1.0) for both autocorrelated-rates (AR) and independent-rates (IR) molecular clock rate model (see Methods). We mainly focused on five calibration strategies. These five calibration strategies consisted of a single calibration at the root with both upper and lower bounds (*root*), fully calibrated root plus one fully-bounded internal node (*calib-1_0.2*), fully calibrated root plus one minimum-bounded internal node (*calib-1_0.2--only_min*), fully calibrated root plus two fully-bounded internal nodes (*calib-2_0.2*), and fully calibrated root plus two minimum-bounded internal nodes (*calib-2_0.2--only_min*).

Under the focal simulation scheme (see Methods: Simulation procedure), the profile mixture model LG+C60+G performed comparably to the traditional LG+G model for divergence time estimation when simulated root ages were set to either 1.0 or 2.0 billion years ago (Ga). However, for root ages ≥3.0 Ga, LG+C60+G obtained significantly more accurate time estimates than LG+G, with the performance gap becoming increasingly pronounced at 4.0 Ga (Fig. 1A). For example, under the *root* calibration strategy, LG+C60+G demonstrated superior time estimation accuracy (mean relative difference ± SD: 0.157 ± 0.061 for AR clock model, 0.139 ± 0.046 for IR clock model) compared to LG+G (AR: 0.193 ± 0.062; IR: 0.197 ± 0.061). Mean relative difference is defined as mean $\left(\frac{|b_i^m - b_i^{true}|}{\max(b_i^m, b_i^{true})}\right)$, where b_i^m and b_i^{true} respectively denote the inferred posterior mean age and the true value for node i. The largest discrepancies in divergence time estimation occurred under the calibration scheme "*calib-2_0.2--only_min*", where LG+C60+G showed ~40% lower error than LG+G (mean relative difference: AR 0.160 vs. 0.232; IR 0.147 vs. 0.259; Fig. 1A). In contrast to divergence time estimates, LG+C60+G provided much more accurate branch-specific substitution rate estimates than LG+G across all tested root ages (Fig. 1B). The error reduction obtained by LG+C60+G compared with LG+G, calculated as $\frac{\text{reldiff}_{\text{LG+G}} - \text{reldiff}_{\text{LG+C60+G}}}{\text{reldiff}_{\text{LG+G}}}$, averaged across all calibration strategies, ranged from 26% to

171 53% under AR and 23% to 55% under IR for root ages of 1.0 to 4.0 Ga. The pattern held true
172 with alternative metrics for comparison with the simulation values (Figs. S1-S3).
173 Next, we explored the patterns for more complex mixture models. We combined structurally
174 constrained substitution models EX2 (exposed/buried sites) and EX3 (highly
175 exposed/intermediate/buried sites) (Le et al. 2008) with the predefined equilibrium frequency
176 C20 to account for site-specific variation in amino acid profiles by 20 categories of equilibrium
177 frequencies. This resulted in 40 and 60 different rate matrices for EX2 and EX3 respectively.
178 When simulation was done under profile-exchangeability mixture models EX2+C20+G and
179 EX3+C20+G, the above patterns became more pronounced (Fig. S4). In contrast to profile
180 mixture models, when alignments were simulated under exchangeability mixture models
181 (EX2+G, EX3+G, and LG4M), LG+G and the true simulation model performed similarly in
182 estimating divergence times and evolutionary rates (Fig. S5).
183 Large sequence divergence may not only be a result of divergence time, but also rapid
184 evolution. To test whether different time estimates are affected by varying evolutionary rates
185 (Fig. S6), we simulated alignments on phylogenies with a fixed root age of 1.0 Ga but different
186 mean absolute substitution rates along the phylogeny (0.25-1.0 substitutions/site/Gyr). This
187 represents evolutionary rates from relatively low, as in most microbes, to very fast, as observed
188 in some pathogens. We would anticipate comparable sequence divergence whether we fixed the
189 root age at 1.0 Ga and increased the average substitution rate (from 0.25 to 1.0
190 substitutions/site/Gyr; Fig. S6), or fixed the substitution rate at 0.25 substitutions/site/Gyr and
191 increased the root age to 4.0 Ga (Fig. 1). Indeed, we observed similar patterns to the focal data
192 when simulating evolution with a high substitution rate. Averaged across all calibration
193 strategies, LG+C60+G displayed error reductions in substitution rate estimate of 27% (AR) and

194 23% (IR) at a substitution rate of 0.25 substitutions/site/Gyr, and 65% (AR) and 61% (IR) at 1.0
195 substitutions/site/Gyr.
196
197 **Performance of profile mixture models in molecular clock dating under varied simulation**
198 **conditions**
199 Next, we evaluated the influence of multiple sources of uncertainties in molecular clock dating
200 on the performance of profile mixture and traditional homogenous models when sequence
201 alignments are simulated under profile mixture models. These involve sequence length, Gamma-
202 distributed across-site rate, birth-death process, variance in lineage-specific rates, and number of
203 sequences (see Note S1 for details).
204 We first investigated how sequence length (ranging from 100 to 2000 aa) affected the
205 accuracy of the estimates using different substitution models (Fig. S7). Although the increase in
206 divergence time estimate accuracy achieved by using LG+C60+G instead of LG+G was minimal
207 for 100-aa alignments, the improvement became increasingly apparent as the alignment length
208 extended to 2000 aa. Across all analyzed calibration strategies, the LG+C60+G model
209 consistently improved the accuracy of time estimation for the AR clock models compared to
210 LG+G. For a simulated root age of 4.0 Ga, this improvement was substantial, reaching 30%,
211 28%, 25%, and 12% for alignment lengths of 2000, 1000, 500, and 100 aa, respectively; the
212 benefit was still significant but less pronounced for a younger root age of 3.0 Ga, with
213 corresponding improvements of 22%, 18%, 15%, and 7%. Both LG+C60+G and LG+G models
214 achieved higher estimation accuracy with longer sequences (i.e., lower values in *y*-axis from 100
215 to 2000 aa), indicating that longer sequences are more informative for more accurate molecular

clock analyses, consistent with the notion that mixture models perform much better on longer alignments for phylogenetic reconstruction (Baños et al. 2024a).

We further adjusted the gamma correction parameter (α) to modulate the variance of the across-site relative rates. Specifically, we set α to 0.5 and 1.5, which scaled the across-site variance to 2-fold ($0.5/0.5^2$) (scheme *Gamma_0.5* in Fig. 2A) and 2/3-fold ($1.5/1.5^2$) (scheme *Gamma_1.5* in Fig. S9A), respectively, compared to the our focal analysis (α=1). These adjustments did not alter the patterns observed in the original analysis. Similarly, when exploring alternative birth-death process parameters, a flat birth-death process (simulation scheme: *bd_flat*) produced results comparable to those from the focal analysis (Fig. 2B). Under complete species sampling (100%), the difference in time estimates between the LG+G and LG+C60+G models became more obvious (scheme *bd_0.4-0.2-1* in Fig. S9B). Moreover, increasing the standard variation of the among-branch sequence absolute substitution rate to twice and three times its default value reduced the differences in divergence time estimation between using LG+G and LG+C60+G, despite the unchanged overall pattern (schemes *sigma_0.38, sigma_0.55* in Figs. 2C, S9F). In addition, although all simulations above were performed on 20-tip trees for simplicity, the analysis was extended to 40-, 60-, 80-, and 100-tip trees, with similar patterns found (Fig. S10).

Profile mixture models display higher robustness to model misspecification

Further, we applied simulations to test the robustness of profile mixture models to misspecifications in calibrations or substitution models. To explore whether mis-specified calibrations would affect estimation under different substitution models when the sequences were simulated under LG+C60+G, we calibrated the nodes by adjusting their ages. These adjustments

consisted of fixed percentage increases (shifting ages toward the past) or decreases (shifting ages toward the present) (Fig. 3). In general, LG+C60+G achieved higher accuracy when calibrations were older than the true divergence times (Fig. 3B). For a root age of 4.0 Ga and across all tested calibration strategies, on average, the relative difference in time estimation compared to true values was: 0.256 (LG+G) versus 0.195 (LG+C60+G) for the AR model, and 0.278 (LG+G) versus 0.185 (LG+C60+G) for the IR model. For root age of 3.0 Ga, these values changed to 0.252 (LG) versus 0.202 (LG+C60+G) for AR, and 0.255 (LG) versus 0.192 (LG+C60+G) for IR (Fig. 3B). Subtle differences between using LG+G and LG+C60+G were observed for root age ≤ 2.0 Ga. However, when calibrations were biased towards younger ages, the LG+G and LG+C60+G models generally yielded comparable accuracy for divergence time and substitution rate estimates, irrespective of the root age, with significant differences arising occasionally.

The above analyses focused on alignments simulated under complex substitution models, particularly LG+C60+G. To determine how these models perform in molecular clock analysis when sequences evolve under simpler substitution models, we performed additional simulations using these models and the previously described settings (varying the simulated root ages from 1.0 to 4.0 Ga (Fig. 3B). These analyses indicate that the profile mixture model LG+C60+G displayed robust time and rate estimation, regardless of the true underlying substitution process. The robustness of LG+C60+G is evident in the following comparisons. When sequences were simulated under the basic LG+G model, LG+C60+G provided estimates of divergence times and absolute substitution rates at least as accurate as those obtained using LG+G. Further, similar results were observed when simulations were conducted with other models with factors not explicitly modeled by LG+C60+G, such as invariant sites (+I) and free rate across-site heterogeneity (+R) (Note S1). These findings suggest that LG+C60+G's robustness was

maintained even when the true evolutionary process involves additional complexities in substitution models.

Comparison between profile mixture models with PMSF (posterior mean site frequency) and bootstrapped-Hessian approximations in molecular clock dating

The computational intensity of the LG+C60+G model often necessitates the PMSF approximation to accelerate phylogenetic reconstruction, mainly by simplifying branch length optimization (Wang et al. 2018). A distinct approximation, previously employed in molecular clock analyses under complex substitution models, uses a bootstrap-based method to approximate the Hessian matrix and is implemented in the software bs_inBV (Wang and Luo 2025). To assess the impact of these two approximations on molecular clock dating under profile mixture models, we compared divergence time and substitution rate estimates derived with each approach against those obtained with the full Hessian matrix calculated without either approximation. We found that both approximations provided accurate estimates of divergence times, no matter which root age was used in simulation (Fig. 3A). For example, across all simulated root ages, compared to LG+C60+G estimated rates, the average differences for AR were 0.036 ± 0.010 (PMSF approximation; *LG+C60+G+PMSF*), 0.071 ± 0.006 (bootstrapped-hessian; *LG+C60+G+bs_inBV*), and 0.073 ± 0.008 (combined use; *LG+C60+G+PMSF+bs_inBV*); for IR, the differences were 1.8% ± 0.4%, 5.9% ± 0.4%, and 6.5% ± 0.5%, respectively. The deviations in time estimates from these approximations, particularly PMSF, were close enough to the stochastic variation observed across independent MCMC runs of the original profile mixture model (*LG+C60+G_run2* in Fig. 4). The estimates of evolutionary rates exhibited slightly greater divergence from the original model with either

approximation (Fig. 3B). Compared to rates estimated under the LG+C60+G model without any approximation, the mean relative differences under both AR and IR models were around 8% for either approximation, or their combination.

Divergence times and substitution rates under mixture models for the ancient symbionts Microsporidia and Rickettsiales: an empirical study

To evaluate mixture models on empirical data, we selected two ancient symbiont lineages with distinct evolutionary significance and accelerated evolutionary rates: Microsporidia, highly reduced, fast-evolving obligate intracellular parasites of animals related to fungi (Capella-Gutiérrez et al. 2012), and Rickettsiales, an alphaproteobacterial lineage renowned as close relatives of mitochondrial ancestors, found in both protists and animals (Martijn et al. 2018; Wang and Luo 2021). Their rapid evolution combined with ancient divergence makes them ideal test cases for our analysis. Note that our primary goal was to assess how complex substitution models influence the molecular clock analysis itself, leading us to test different settings rather than selecting a single optimal configuration (e.g., partitions, clock model).

For both datasets, we partitioned the alignments into three subsets (see Methods) according to a Gaussian mixture model-based clustering according to the logarithm of the estimated substitution rates of each gene, and selected the best-fitting substitution models for each partition under both site-homogeneous and mixture substitution models respectively. In all cases, profile mixture models were associated with the lowest AIC values, with LG+Cxx+F+I+R4 (where xx = 40, 50, 60) being the most frequently selected (Data S3-S4). Microsporidia do not have fossil records. Thus, we applied different calibrations to only the root node based on previous estimate of the occurrence time of its LCA (last common ancestor) (Note S2). We found that the time

308 estimates for Microsporidia lineages were largely consistent between the best-fitting profile
309 mixture and homogenous models (Fig. 5). The LCA of Microsporidia was estimated to occur at
310 approximately 0.5 Ga and 0.65 Ga under the AR and IR clock models, respectively (Fig. 5A). In
311 contrast to the similar divergence time estimates between homogeneous and mixture models, the
312 absolute substitution rate was at least 1/3 higher under the mixture model (Fig. 5B). All
313 microsporidian lineages experienced accelerated sequence changes compared to the outgroup
314 lineages. While the highest substitution rate was predicted for the branch leading to the LCA of
315 Microsporidia under the IR clock model, the AR clock model instead predicted the highest rate
316 on the branch leading to *Enterocytozoon*, a major microsporidian pathogen in human and
317 animals. The above pattern was consistent with alternative calibrations and different numbers of
318 partitions (Fig. S11).

319 Rickettsiales also lack direct fossil evidence; however, due to their close phylogenetic
320 relationship with mitochondria, we employed the mitochondrial endosymbiosis dating strategy as
321 used in our prior study (Wang and Luo 2021). This approach allows estimating Rickettsiales'
322 evolutionary timeline by incorporating eukaryotic fossils, based on the evolutionary scenario that
323 mitochondria evolved from a lineage closely related to Alphaproteobacteria (and thus
324 Rickettsiales). Particularly, under the AR clock model, the time estimates of Rickettsiales
325 lineages under the mixture substitution model differed by an average of >20% compared to those
326 from the homogeneous alternatives (Fig. 6A). In general, the ages of most lineages were
327 estimated to be older under the profile mixture model (Fig. 6A). For example, under the AR
328 clock model, the LCA of Rickettsiales was estimated to occur at 2.67 Ga (95% HPD: 2.97-2.35
329 Ga) by the mixture model, whereas the homogeneous substitution model estimated it at 2.26 Ga
330 (95% HPD: 2.57-1.94 Ga). Similarly, eukaryotes were estimated to originate at 2.29 Ga (95%

331 HPD: 2.57-1.98 Ga) under the mixture model, compared to 1.74 Ga (95% HPD: 1.98-1.50 Ga)
332 under the homogeneous model (Fig. 6A). The above pattern holds under alternative calibration
333 settings (Fig. S12). With a single partition (Fig. S12A, S12B), the difference in time estimates
334 between different substitution models was less obvious. In all calibration schemes, the
335 substitution rates under the mixture models were roughly 1.5 to 2 times those estimated under
336 homogeneous models (Fig. 6B).

337 A recent study (Baños et al. 2024a) suggests that the empirical frequency estimated from the
338 alignment (+F) might mislead phylogenetic reconstruction for profile mixture models. As an
339 additional analysis, we re-performed molecular clock dating with the best-fitting mixture model
340 without the empirical frequency option. Similar patterns were found (Figs. S13-S14).

341

342 **Discussion**

343 **Approximate likelihood dating with complex models**

344 The present study introduces phyloHessian, enabling Hessian calculation under diverse mixture
345 and non-mixture amino acid substitution models for molecular clock dating using MCMCtree's
346 approximate likelihood method (Thorne et al. 1998; dos Reis and Yang 2011). A key question is
347 whether MCMCtree's approximate likelihood method performs equally well as exact likelihood
348 under mixture substitution models in terms of divergence time estimation. While previous
349 research suggests that this holds true for homogeneous models (dos Reis and Yang 2011), a full
350 comparison is impossible due to the lack of software directly implementing complex substitution
351 models. Therefore, we compared the phylogenetic likelihoods calculated by the approximate and
352 exact likelihood under LG+C60+G using phyloHessian. The resulting log-likelihood curves were
353 similar (Fig. S15). This suggests that MCMCtree’s approximate likelihood dating roughly well

approximates the log-likelihood surface under mixture models. A more thorough examination in time estimates, instead of the phylogenetic likelihood, is necessary to fully address the question in future.

It was when we were finishing writing this manuscript that another computational software IQ2MC (Demotte et al. 2025) implemented in the newest version of IQ-Tree (v3.0.1) (Wong et al. 2025), became available on a preprint server. IQ2MC leverages the highly efficient computation of IQ-Tree's phylogenetic library to estimate the Hessian matrix under not only amino acid, but nucleic acid and partitioned models, unlike PhyloHessian which is restricted to amino acid models. IQ2MC was roughly 3X and 10X times faster than CODEML and PhyloHessian respectively (Data S5). However, our software, phyloHessian, offers more flexible settings in molecular dating by alternative Hessian calculation algorithms and interactions with other software like PhyML and bs_inBV. Its code for phylogenetic likelihood calculation under complex substitution models would be a valuable resource for phylogenetic software development within the Julia ecosystem.

Other methods, such as mcmc-date (Mahendrarajah et al. 2023), also accelerate molecular dating analyses by approximating the phylogenetic likelihood. It is in concept similar to bs_inBV (Wang and Luo 2025), but operates in a Bayesian framework. A key advantage of bs_inBV and mcmc-date over PhyloHessian and IQ2MC is their ability to accommodate substitution models not yet available in specific phylogenetics software, such as FunDi (Gaston et al. 2011) or CAT-PMSF (Szánthó et al. 2023). However, these approximations via bootstrap or MCMC sampling may sacrifice computational speed and accuracy, to some extent (Fig. 4).

Deep-time molecular clock dating: the influence of complex mixture models

377 Our extensive simulations demonstrate that the advantage of using the complex mixture
378 substitution model for divergence time estimation becomes increasingly significant as
379 evolutionary timescales deepen (Figs. 1-4). Particularly, when the root age was earlier than 2.0-
380 1.0 Ga, the difference became obvious. This highlights the importance of substitution models
381 that better fit the data for dating deep-time phylogenies. On the other side, the substitution rate
382 was almost always estimated higher and more accurate under mixture models than under
383 homogeneous models. The underlying logic is straightforward. Mixture models typically
384 estimate longer branch lengths than homogeneous models (Venditti et al. 2008; Moody et al.
385 2022; Wang and Luo 2025). Because branch length equals substitution rate multiplied by time
386 (dos Reis et al. 2016; Guindon 2020), the underestimated branch lengths under simpler
387 substitution models must be attributable to changes in either the rate, the time, or both. Given
388 that molecular clock analyses typically constrain the time estimates through fossil-based
389 calibrations, the increased branch lengths under mixture models primarily result in higher rate
390 estimates, with a smaller impact on the time estimates themselves.

391 The case in empirical data analysis is more complex. Unlike simulations with known
392 parameters, empirical data could arise from processes so complex that even the best-fitting
393 models are inadequate. This notion is supported by the frequent selection of the most highly
394 parameterized models as the best fit, implying that the complexity of the real process the
395 sequence is generated far exceeds the capabilities of our current models. Consequently, they
396 could make little difference in the time estimates compared to homogeneous models, because
397 both might be an oversimplification of the true model. In such cases, it would be worth applying
398 models like UDM (Schrempf et al. 2020) with more profile mixtures than C60, or the GTRpmix
399 exchangeability matrix and its derived models (Baños et al. 2024b). While these models have

been integrated into phyloHessian, a comprehensive evaluation of their applications in molecular dating analysis on empirical data remains a task for future research.

While simulations suggest that homogeneous models likely overestimate divergence times in scenarios mimicking deep-dating evolution (Figs. S2-S3), the Rickettsiales data reveals a contrasting pattern where mixture models led to older age estimates (Fig. 6). We suspect that one important factor lies in the calibration settings. In simulations, we have complete control over the calibration settings. Real-world data introduces complexities absent in simulation because the available calibrations are not uniform in their quality. While some might accurately reflect the true age of the node, others could be overly young, overly old, or simply uninformative. This complex interplay between substitution models and the quality and number of time calibrations, along with other variables such as the clock model and sequence length, makes the impact of mixture substitution models on posterior age estimates case dependent.

In summary, while the present study indicates significant impacts of complex profile mixture substitution models on divergence time inferences for deep-time or fast-evolving lineages, the importance of these models on time estimates cannot be overstated; their effect likely requires consideration on a dataset-specific basis. Therefore, we recommend rigorous molecular clock analyses employing different substitution models to assess their impact on divergence time and substitution rate estimates in future research.

Implications for the evolution of ancient symbionts

For empirical datasets of Microsporidia and Rickettsiales, profile mixture models, which display much higher model fit to data than their homogeneous counterparts, produced significantly higher substitution rates than estimated by homogeneous models (Figs. 5B, 6B). The difference

423 suggests that bacterial pathogens/symbionts could experience relatively high evolutionary rates
424 up to 1.0 or 2.0 Gyr. This hints that factors like elevated mutation rates, selection, or persistently
425 reduced effective population sizes, which are often responsible for rapid sequence divergence,
426 can have more profound and long-lasting effects on the evolutionary trajectory of symbiotic
427 microorganisms than previously thought. Studying the substitution rates will also improve our
428 understanding of microbial lifestyle evolution.

429 One of the merits of the mitochondria symbiosis-based dating method for Rickettsiales
430 evolution is that it co-estimates the divergence times for both Rickettsiales and their eukaryotic
431 hosts with a unified dataset (Wang and Luo 2021). Notably, under profile mixture models, the
432 origin time of Rickettsiales was estimated at 2.5-2.0 Ga, roughly 0.5 Gyr earlier than that of the
433 last eukaryotic common ancestor (LECA) using homogeneous models estimated in both the
434 current (Fig. 6) and previous studies (Parfrey et al. 2011; Betts et al. 2018; Wang and Luo 2021;
435 Moody et al. 2024). Hence, if the revised time estimate is accurate, it implies that the
436 Rickettsiales' LCA existed before LECA emerged. This interpretation is consistent with recent
437 comparative genomics findings suggesting that early Rickettsiales initially adopted a free-living
438 or non-obligate symbiotic lifestyle (Castelli et al. 2024), with subsequent evolutionary events
439 leading to independent colonizations of eukaryotic hosts—first protists, then animals (Castelli et
440 al. 2016; Wang and Luo 2021). Alternatively, it is possible that early Rickettsiales were
441 associated with unknown host organisms belonging to eukaryotic stem groups (i.e., extinct or
442 unknown lineages that diverged before LECA but after the split with the closest living relatives
443 of eukaryotes). This hypothesis points to previously unrecognized interactions between stem-
444 group eukaryotes and prokaryotes, and hints at complex prokaryotic engagement as a
445 foundational trait of eukaryotic ancestors, offering insights into the origin of eukaryotes.

Methods

Mixture substitution model

In a mixture model, each category consists of a rate matrix M, an equilibrium frequency vector F, and a relative substitution rate R. Additionally, the computation of the likelihood necessitates the inclusion of branch lengths θ, and the tree topology τ. Let the full parameter vector be denoted by $\mathbf{\Theta} = (R, F, V, \theta, \tau)$. Denote by $\mathcal{R}$, $\mathcal{F}$, and $\mathcal{V}$ the number of categories of exchangeability matrices, equilibrium frequencies, and relative substitution rates, respectively. Denote by $\mathcal{S}$ the number of alignment sites. Hence, the full likelihood of the phylogeny is given as

$$P(D|\mathbf{\Theta}) = \prod_{s=1}^{\mathcal{S}} \sum_{r=1}^{\mathcal{R}} \sum_{f=1}^{\mathcal{F}} \sum_{v=1}^{\mathcal{V}} P\left(D_s \middle| R_r, F_f, V_v, \theta, \tau\right) P\left(R_r, F_f, V_v\right) I(r, f, v)\,, \qquad (1)$$

where $I(m, f, v)$ is the joint indicator function that takes value 1 if the combination M_m, F_f, V_v is defined in the model and 0 otherwise (Guindon 2019). In most cases, the exchangeability matrices, equilibrium frequencies, and relative rates are considered independent, hence $I(r, f, v) = \mathbb{1}_{1 \le r \le \mathcal{R}} \cdot \mathbb{1}_{1 \le f \le \mathcal{F}} \cdot \mathbb{1}_{1 \le v \le \mathcal{V}}$. An exception is the LG4M and LG4X substitution models (Le et al. 2012) whose equilibrium frequencies and relative rates are linked such that $I(r, f, v) = \mathbb{1}_{1 \le r \le \mathcal{R}} \cdot \mathbb{1}_{1 \le f \le \mathcal{F}} \cdot \mathbb{1}_{1 \le v \le \mathcal{V}} \cdot \delta_{f,v}$ where $\delta_{f,v}$ is the Kronecker delta which equals 1 if $f = v$ and 0 otherwise. Most commonly, it is assumed that the different mixture components are independent such that $P\left(R_r, F_f, V_v\right) = \frac{P(R_r)P\left(F_f\right)P(V_v)}{\sum_{r,f,r} P(R_r)P\left(F_f\right)P(V_v)I(r,f,v)}$.

Numerical calculation of the Hessian matrix

Two primary numerical methods are commonly employed for computing the Hessian matrix. Let $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ indicate the vector of the MLES of branch lengths, and $\ell(\hat{\theta})$ be the log-likelihood of the phylogeny. The outer product of scores (OPS) method (Seo et al. 2004; dos Reis and Yang 2011) estimates the Hessian matrix indirectly using the first derivatives (scores) of the log-likelihood to sum the contributions of the log-likelihood at each site:

$$H(\hat{\theta}) = -\sum_{s=1}^{S} \left(\frac{\partial}{\partial \theta} \ell_s(\hat{\theta}) \frac{\partial}{\partial \theta^T} \ell_s(\hat{\theta}) \right). \tag{2}$$

Each component of the score vector $\frac{\partial}{\partial \theta} \ell_s(\hat{\theta})$ is numerically calculated by the central finite difference method as

$$\frac{\partial \ell_s}{\partial \theta_i} \approx \frac{\ell_s(\hat{\theta} + he_i) - \ell_s(\hat{\theta} - he_j)}{2h} + O(h^2),$$

where e_i and e_j are the i-th and j-th standard basis vectors respectively, and h is a small step size, typically the cube or square root of machine precision.

The second way is by a direct second-order derivative method. It calculates $H(\hat{\theta})$ via finite differences of the log-likelihood function by

$$H(\hat{\theta}) = \sum_{s=1}^{S} \frac{\partial^2}{\partial \theta \partial \theta^T} \ell_s(\hat{\theta}), \tag{3}$$

where diagonal elements and off-diagonal elements are respectively calculated by the central finite difference method as

$$H_{ii}(\hat{\theta}) \approx \frac{\ell(\hat{\theta} + he_i) - 2\ell(\hat{\theta}) + \ell(\hat{\theta} - he_i)}{h^2},$$

$$H_{ij}(\hat{\theta}) \approx \frac{\ell(\hat{\theta} + he_i + he_j) - \ell(\hat{\theta} + he_i - he_j) - \ell(\hat{\theta} - he_i + he_j) + \ell(\hat{\theta} - he_i - he_j)}{4h^2}.$$

Under regularity conditions, and asymptotically as the number of sites $\mathcal{S}$ increases, both methods converge to the true Hessian:

$$\sum_{s=1}^{\mathcal{S}} \frac{\partial^2}{\partial\theta\partial\theta^T} \ell_s(\hat{\theta}) \simeq -\sum_{s=1}^{\mathcal{S}} \left(\frac{\partial}{\partial\theta} \ell_s(\hat{\theta}) \frac{\partial}{\partial\theta^T} \ell_s(\hat{\theta}) \right). \tag{4}$$

In both methods described as above, the error term is of order $\mathcal{O}(h^2)$. To further improve the numerical accuracy of the second-order derivative method, particularly when dealing with very short branches, we derive the central finite difference calculation of the Hessian at $\mathcal{O}(h^4)$. For the OPS method, this is simply

$$\frac{\partial \ell}{\partial \theta_i} \approx \frac{-\ell(\hat{\theta} + 2he_i) + 8\ell(\hat{\theta} + he_i) - 8\ell(\hat{\theta} - he_i) + \ell(\hat{\theta} - 2he_i)}{12h} + \mathcal{O}(h^4) \tag{5}$$

followed by summing the contributions of the log-likelihood at each site according to Eq. (2).

As to the direct second-order derivative method, the multivariate Taylor expansion of $\ell(\theta)$ around the point (θ_i, θ_j) is given by

$$\ell(\hat{\theta} + e_i a + e_j b) = \underbrace{\ell + a\frac{\partial \ell}{\partial \theta_i} + b\frac{\partial \ell}{\partial \theta_j} + \sum_{k=2}^{5} \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} a^m b^{k-m} \frac{\partial^k \ell}{\partial {\theta_i}^m \partial {\theta_j}^{k-m}}}_{\text{Term 1}} + \mathcal{O}(a^6, b^6).$$

Evaluating $\ell(\hat{\theta})$ at $1 + 4 \times 4 = 17$ symmetric points $(\hat{\theta}_i, \hat{\theta}_j), (\hat{\theta}_i \pm h, \hat{\theta}_j \pm h), (\hat{\theta}_i \pm 2h, \hat{\theta}_j \pm 2h)$ and their combinations, while holding all other variables $\hat{\theta}_k$ for $k \neq i, j$ constant, we construct $Ax = B$. In this system, $A \in R^{21\times17}$ is the matrix of Taylor coefficients for derivatives up to the 5th order, $x \in R^{17}$ is the vector of weights for the function evaluations, and $B \in R^{17}$ is the target vector encoding the desired derivative and cancellation of higher-order terms (Note S3). The row size of A is 21 because there are 21 items in Term 1 to eliminate to achieve an error order of $\mathcal{O}(h^4)$. The system is underdetermined as its rank is 15, smaller than the number of equations which is 17, leading to infinitely many solutions. To find out the solution that

minimizes $||x||_F^2$, we use the formula $x = A^T(AA^T)^{-1}B$, where $A^T(AA^T)^{-1}$ is also known as the right pseudo-inverse of A. Denote $\ell^{(k,j)} = \ell\left(\hat{\theta} + khe_i + lhe_j\right)$. These give

$$H_{ii}(\hat{\theta}) \approx \frac{1}{h^2}\left(-\frac{5}{4}\Delta^{(0,0)} + \frac{1}{3}\Delta^{(1,1)} - \frac{1}{48}\Delta^{(2,2)} + \frac{12}{169}\Delta^{(1,2)} - \frac{12}{169}\Delta^{(2,1)}\right), \quad (6)$$

$$H_{ij}(\hat{\theta}) \approx \frac{1}{600h^2}\left(74\Delta^{(1,1)} + 44\Delta^{(2,2)} - 63\Delta^{(1,2)} + 63\Delta^{(2,1)}\right), \quad (7)$$

where $\begin{cases} \Delta^{(1,1)} = \ell^{(1,1)} - \ell^{(1,-1)} - \ell^{(-1,1)} + \ell^{(-1,-1)} \\ \Delta^{(2,2)} = -\ell^{(2,2)} + \ell^{(2,-2)} + \ell^{(-2,2)} - \ell^{(-2,-2)} \\ \Delta^{(1,2)} = \ell^{(1,2)} - \ell^{(1,-2)} - \ell^{(-1,2)} + \ell^{(-1,-2)} \\ \Delta^{(2,1)} = \ell^{(2,1)} - \ell^{(2,-1)} - \ell^{(-2,1)} + \ell^{(-2,-1)} \end{cases}$.

More details on math are given in Note S3.

In phyloHessian, the default algorithm is the OPS method, which is employed by CODEML in PAML v4.3+. Alternatively, users can specify the second-order derivative method, as used in PAML 4.1 and earlier versions. Users can also specify the order of the error term, allowing balancing computational accuracy with speed based on their needs.

Simulation procedure

We evaluated the performance of different substitution models using simulated data that mimics deep-time evolution (Fig. S16). In our focal simulation scheme, for each substitution model used in sequence evolution simulation (details provided below), we generated 30 timetrees each with 20 tips, using the R package TreeSim (Stadler 2011). Following our previous work (Wang and Luo 2025), timetrees were created under a birth-death model with birth and death rates of 0.4 and 0.2 lineages per 0.1 Gyr as estimated in a prior study of prokaryotic lineage diversification (Scholl and Wiens 2016), and with a taxon sampling proportion of 10% (see Note S1 for alternative settings). Simulated trees were generated with true root ages of 1.0, 2.0, 3.0, and 4.0 Ga to represent a range of evolutionary timescales approaching the age of Earth (~4.5 Ga).

For each timetree, branch-specific substitution rates were sampled under both the independent rates (IR) and autocorrelated rates (AR) model. For the IR mode, following our previous study (Wang and Luo 2021), the rate follows lognormal distribution with parameters $\mu_{\mathrm{IR}} = -3.7$ and $\sigma_{\mathrm{IR}} = 0.2$, such that the mean of rate is $\mathrm{e}^{\mu_{\mathrm{IR}}+\frac{\sigma_{\mathrm{IR}}^2}{2}} \approx 0.025$ substitutions/site/0.1 Gyr, with a standard variation of $\sqrt{\left(e^{\sigma_{\mathrm{IR}}^2}-1\right)e^{\left(2\mu_{\mathrm{IR}}+\sigma_{\mathrm{IR}}^2\right)}} \approx 0.005$ substitutions/site/0.1 Gyr. This roughly corresponds to empirical observations that tip-to-root distances in the bacterial tree of life, constructed with universally conserved orthologs, are roughly 1.0 amino acid substitutions/site (Moody et al. 2022; Wang and Luo 2025), assuming an origin of early life at around 4.0 Ga (Battistuzzi and Hedges 2009; Marin et al. 2017; Betts et al. 2018; Moody et al. 2024). The methodology used to determine the simulation parameters for the AR model is described in detail in the next section.

Alignments of 300 amino acids (aa) were simulated using AliSim (Ly-Trong et al. 2022) according to the timetree generated by simulation, under various substitution models. In the focal analysis, we employed the LG+C60+G4{1.0} model. This indicates LG substitution model plus C60 profile mixture models (60 empirically derived amino acid site frequency profiles) (Quang et al. 2008) and across-site rate variation under a discrete gamma distribution with four relative rate categories parameterized by α=1.0, hence Gamma(1,1), meaning that the mean and variance of the across-site relative rates are $\frac{1}{1} = 1$ and $\frac{1}{1^2} = 1$ respectively (Yang 1994). Further, we varied several simulation parameters, including sequence length, number of taxa, calibration strategies, birth-death process parameters, across-site rate heterogeneity, molecular clock settings, and model misspecifications. These are detailed in Note S1.

Determining the parameters for simulation under the AR clock rate model

Denote the absolute rate of sequence evolution (measured in substitutions per site per time unit) under IR and AR respectively as the vectors $X_{\mathrm{IR}} = [X_{1,\mathrm{IR}}, \dots, X_{n,\mathrm{IR}}]$ and $X_{\mathrm{AR}} = [X_{1,\mathrm{AR}}, \dots, X_{n,\mathrm{AR}}]$ where n represents the number of branches. Their log-transformed values are given by $Y_{\mathrm{IR}} = [Y_{1,\mathrm{IR}}, \dots, Y_{n,\mathrm{IR}}]$ and $Y_{\mathrm{AR}} = [Y_{1,\mathrm{AR}}, \dots, Y_{n,\mathrm{AR}}]$. To make simulations under IR and AR models more comparable, we require that the expectations of the sample mean and sample variance of the log-transformed rate under the two models to be equal, respectively, given the same phylogeny. Mathematically, this requires that

$$\mathrm{E}(\bar{Y}_{\mathrm{AR}}) = \mathrm{E}(\bar{Y}_{\mathrm{IR}}), \quad \mathrm{E}(s^2_{\mathrm{AR}}) = \mathrm{E}(s^2_{\mathrm{IR}}) \tag{8}$$

where the sample variances of the log-transformed rates are denoted by s^2_{AR} and s^2_{IR}.

For the IR model, it is assumed that the rate for each branch, denoted $X_{\mathrm{IR}} = [X_{1,\mathrm{IR}}, \dots, X_{n,\mathrm{IR}}]$, is independently and identically (*i.i.d.*) distributed according to a lognormal distribution with parameters μ_{IR} and σ^2_{IR}, such that the rate for each branch can be expressed as $X_{i,\mathrm{IR}} \sim \mathrm{Lognormal}(\mu_{\mathrm{IR}}, \sigma^2_{\mathrm{IR}})$. In other words, we have

$$\mathrm{E}(\bar{Y}_{\mathrm{IR}}) = \mu_{\mathrm{IR}}, \quad \mathrm{E}(s^2_{\mathrm{IR}}) = \sigma^2_{\mathrm{IR}}.$$

The AR model in molecular clock dating, also known as the geometric Brownian (GBM) model, assumes that the lineage-specific rate follows a geometric Brownian motion with a drift parameter of zero. It is thus parameterized by two parameters, μ_{AR}, the ancestral rate of the log-transformed rate, and σ^2_{AR}, which determines the volatility of the rate changes over time. Accordingly, it can be shown that the log-transformed rate under AR model after time t follows $\mathcal{N}\left(\mu_{\mathrm{AR}} - \frac{1}{2}t\sigma^2_{\mathrm{AR}}, t\sigma^2_{\mathrm{AR}}\right)$ (Rannala and Yang 2007; Guindon 2020). Denote $C = [c_{ij}]_{n\times n}$ the branch-wise covariance matrix where n is the number of branches in a rooted tree (note that this

is different from what is so-called phylogenetic covariance matrix which is related to only the tips [Fig. S17]). Further, the log-transformed rates Y_{AR} follow the following multivariate distribution

$$Y_{\mathrm{AR}} \sim \mathcal{N}\left(\left[\mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{11} \cdots \mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{nn}\right]^{\mathrm{T}}, \sigma_{\mathrm{AR}}^2 C\right).$$

Thus, it is easy to see that

$$\mathrm{E}\left(Y_{i,\mathrm{AR}}\right) = \mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{ii}. \tag{9}$$

The expectation of the sample variance s_{AR}^2 of log-transformed rate Y_{AR} across all n branches is calculated as

$$\mathrm{E}(s_{AR}^2) = \frac{1}{n-1}\mathrm{E}\left(\sum\nolimits_{i=1}\left(Y_{i,\mathrm{AR}} - \frac{1}{n}\sum\nolimits_{j=1} Y_{j,\mathrm{AR}}\right)^2\right).$$

Expanding the right-hand side, we obtain

$$\mathrm{E}\left[\sum_{i=1}^{n}\left(Y_{i,\mathrm{AR}} - \frac{1}{n}\sum_{j=1}^{n} Y_{j,\mathrm{AR}}\right)^2\right]$$

$$= \sum_{i=1}^{n}\mathrm{E}\left(Y_{i,\mathrm{AR}}^2\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} 2\mathrm{E}\left(Y_{i,\mathrm{AR}}Y_{j,\mathrm{AR}}\right) - \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathrm{E}\left(Y_{j,\mathrm{AR}}Y_{k,\mathrm{AR}}\right)\right)$$

$$= \sum_{i=1}^{n}\mathrm{E}\left(Y_{i,\mathrm{AR}}^2\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left(Y_{i,\mathrm{AR}}Y_{j,\mathrm{AR}}\right). \tag{10}$$

Further, note the following relationship

$$\mathrm{E}\left(Y_{i,\mathrm{AR}}^2\right) = \mathrm{Var}\left(Y_{i,\mathrm{AR}}\right) + \left[\mathrm{E}\left(Y_{i,\mathrm{AR}}\right)\right]^2 = c_{ii}\sigma_{\mathrm{AR}}^2 + \left(\mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{ii}\right)^2,$$

$$\mathrm{E}\left(Y_{i,\mathrm{AR}}Y_{j,\mathrm{AR}}\right) = \mathrm{Cov}\left(Y_{i,\mathrm{AR}}, Y_{j,\mathrm{AR}}\right) + \mathrm{E}\left(Y_{i,\mathrm{AR}}\right)\mathrm{E}\left(Y_{j,\mathrm{AR}}\right)$$

$$= \sigma_{\mathrm{AR}}^2 c_{ij} + \left(\mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{ii}\right)\left(\mu_{\mathrm{AR}} - \frac{1}{2}\sigma_{\mathrm{AR}}^2 c_{jj}\right).$$

Introducing the above into Eq. (10), we have

$$\sum_{i=1}^{n} \mathrm{E}\left(Y_{i,\mathrm{AR}}^2\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\left(Y_{i,\mathrm{AR}}Y_{j,\mathrm{AR}}\right)$$

$$= \sum\nolimits_i \left[c_{ii}\sigma_{\mathrm{AR}}^2 + \left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)^2\right] - \frac{1}{n}\sum\nolimits_{i,j}\left[c_{ij}\sigma_{\mathrm{AR}}^2 + \left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)\left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{jj}\sigma_{\mathrm{AR}}^2\right)\right]$$

$$= \underbrace{\sigma_{\mathrm{AR}}^2\left[\sum\nolimits_i c_{ii} - \frac{1}{n}\sum\nolimits_{i,j} c_{ij}\right]}_{\text{Term 1}}$$

$$+ \underbrace{\left[\sum\nolimits_i \left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)^2 - \frac{1}{n}\sum\nolimits_{i,j}\left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)\left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{jj}\sigma_{\mathrm{AR}}^2\right)\right]}_{\text{Term 2}}$$

Substituting $\sum_{\mathrm{i}} c_{ii} = \mathrm{tr}(C)$ and $\sum_{i,j} c_{ij} = \frac{1}{n}\mathbf{1}'C\mathbf{1}$ into Term 1, we can rewrite it as

$$\sigma_{\mathrm{AR}}^2\left[\sum\nolimits_i c_{ii} - \frac{1}{n}\sum\nolimits_{i,j} c_{ij}\right] = \sigma_{\mathrm{AR}}^2\left[\mathrm{tr}(C) - \frac{1}{n}\mathbf{1}'C\mathbf{1}\right].$$

For Term 2, noting that terms involving μ_{AR} cancels out, we have

$$\sum\nolimits_i \left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)^2 - \frac{1}{n}\sum\nolimits_{i,j}\left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{ii}\sigma_{\mathrm{AR}}^2\right)\left(\mu_{\mathrm{AR}} - \frac{1}{2}c_{jj}\sigma_{\mathrm{AR}}^2\right)$$

$$= \left(n\mu_{\mathrm{AR}}^2 - \mu_{\mathrm{AR}}\sigma_{\mathrm{AR}}^2\mathrm{tr}(C) + \frac{1}{4}\sum\nolimits_i c_{ii}^2\sigma_{\mathrm{AR}}^4\right) + \left(-n\mu_{\mathrm{AR}}^2 + \mu_{\mathrm{AR}}\sigma_{\mathrm{AR}}^2\mathrm{tr}(C) - \frac{1}{4n}\sum\nolimits_{i,j} c_{ii}c_{jj}\sigma_{\mathrm{AR}}^4\right)$$

$$= \frac{\sigma_{\mathrm{AR}}^4}{4}\left[\sum\nolimits_i c_{ii}^2 - \frac{1}{n}\left(\mathrm{tr}(C)\right)^2\right].$$

Hence,

$$\mathrm{E}(s_{\mathrm{AR}}^2) = \frac{\sigma_{\mathrm{AR}}^2}{n-1}\left[\left(\mathrm{tr}(C) - \frac{1}{n}\mathbf{1}'C\mathbf{1}\right) + \frac{\sigma_{\mathrm{AR}}^2}{4}\left(\sum\nolimits_i c_{ii}^2 - \frac{1}{n}\left(\mathrm{tr}(C)\right)^2\right)\right].$$

Because $\mathrm{E}(s_{\mathrm{IR}}^2) = \sigma_{\mathrm{IR}}^2$, equating $\mathrm{E}(s_{\mathrm{IR}}^2)$ with $\mathrm{E}(s_{\mathrm{AR}}^2)$ and rearranging terms yields the quadratic equation for σ_{AR}^2

$$\frac{1}{4}\left[\sum_i c_{ii}^2 - \frac{1}{n}\left(\mathrm{tr}(C)\right)^2\right]\sigma_{\mathrm{AR}}^4 + \left[\mathrm{tr}(C) - \frac{1}{n}\mathbf{1}'C\mathbf{1}\right]\sigma_{\mathrm{AR}}^2 - (n-1)\sigma_{\mathrm{IR}}^2 = 0. \quad (11)$$

Discarding the negative root, along with solving Eq. (9), we ultimately get

$$\hat{\sigma}_{\mathrm{AR}}^2 = \frac{2(n-1)\sigma_{\mathrm{IR}}^2}{\sqrt{a^2 + 4b(n-1)\sigma_{\mathrm{IR}}^2} + a}, \quad \hat{\mu}_{\mathrm{AR}} = \mu_{\mathrm{IR}} - \frac{\hat{\sigma}_{\mathrm{AR}}^2 \mathrm{tr}(C)}{2n} \quad (12)$$

where $a = \mathrm{tr}(C) - \frac{1}{n}\mathbf{1}'C\mathbf{1}$, $b = \frac{1}{4}\left[\sum_i c_{ii}^2 - \frac{1}{n}\left(\mathrm{tr}(C)\right)^2\right]$, and $\hat{\sigma}_{\mathrm{AR}}^2$ and $\hat{\mu}_{\mathrm{AR}}$ are the values of the parameters for the AR model that satisfy Eq. (8) given the parameters for the IR model (σ_{IR}^2 and μ_{IR}).

Calibrations used in molecular dating analysis on simulated datasets

Molecular dating analysis were conducted using MCMCtree v4.10.7 (see also Note S3) on phylogenies simulated with procedures described above under varying calibration strategies: a root-only calibration, or root calibration plus one internal calibration, chosen as the node at the median age of all nodes in the true timetree, or two internal calibrations which targeted the 1/3 and 2/3 age quantiles. Time priors for calibrated nodes were set to be uniform within the interval [true_age - (true_age/5), true_age + (true_age/5)], with soft bounds (MCMCtree's default: 2.5% probability of age outside bounds). Other information about MCMCtree analysis can be found in Note S3.

Comparing the performance of different substitution models on simulated datasets

We evaluated the performance of various substitution models mainly using the relative difference (reldiff) that compared the posterior mean divergence times and substitution rates estimated by MCMCtree (using each model) to the true values used in simulations. Relative difference:

calculated as $mean\left(\frac{|b_i^m - b_i^{true}|}{\max(b_i^m, b_i^{true})}\right)$, where b_i^m and b_i^{true} denote the posterior mean divergence times or branch-specific substitution rates inferred by MCMCtree, and those used in simulation (true values), respectively.

Empirical data analysis

Genomes of Microsporidia and Rickettsiales were downloaded from MicrosporadiaDB (Aurrecoechea et al. 2011) and GenBank (last accessed July 2024), respectively. Sequences of five additional Proteobacteria as the outgroup were retrieved from our prior study (Wang and Luo 2025). TreeCluster v1.0.4 (Balaban et al. 2019) was used to select representative taxa with a cutoff of phylogenetic depth (which take phylogenetic distance and topology into account to cluster taxa into subclades) of 0.5 and 2.5 respectively for Microsporidia (yielding 27 members) and Rickettsiales (22 members). Genes used to reconstruct phylogeny and molecular dating for Microsporidia were based on 27 orthologs identified by OrthoFinder v3.0.1 (Emms and Kelly 2019) present in all members allowing ≤2 copies per genome, while for Rickettsiales we adopted the mitochondrial endosymbiosis-based strategy (Wang and Luo 2021) by using 24 mitochondria-encoded genes conserved across Alphaproteobacteria (Wang and Wu 2015) (see Data availability). For both datasets, we partitioned the alignments into one, two, or three subsets by fitting the logarithm of the average substitution rates estimated by CODEML into a Gaussian mixture model. We selected the best-fitting substitution models for each partition under both site-homogeneous and site-heterogeneous models according to AIC (Note S3). Calibration information along with alternatives is available in Data S6.

Data availability

phyloHessian is available at https://github.com/evolbeginner/phyloHessianWrapper. Other codes and the scripts (Goto et al. 2010) generating them, are made available at https://figshare.com/s/96bc117f18051f9642ce.

Acknowledgements

We thank Mario dos Reis, Jianhao Lv, Gergely Szöllősi, Ziheng Yang, and Tianqi Zhu for discussion. We also thank Bui Quang Minh for kindly sharing with us their manuscript, and Margaret Ip for her support. The work is supported by Hong Kong Research Grants Council (RGC) General Research Fund (GRF) (14112024), Natural Science Foundation of China (32400493, 42293294), CUHK Direct Grant (4054912), and the Alliance of International Science Organizations (ANSO) (ANSO-CR-PP-2022-09).

References

Aurrecoechea C., Barreto A., Brestelli J., Brunk B.P., Caler E.V., Fischer S., Gajria B., Gao X., Gingle A., Grant G., Harb O.S., Heiges M., Iodice J., Kissinger J.C., Kraemer E.T., Li W., Nayak V., Pennington C., Pinney D.F., Pitts B., Roos D.S., Srinivasamoorthy G., Stoeckert C.J., Treatman C., Wang H. 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. Nucleic Acids Res. 39:D612-619.

Balaban M., Moshiri N., Mai U., Jia X., Mirarab S. 2019. TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS ONE. 14:e0221068.

Baños H., Susko E., Roger A.J. 2024a. Is Over-parameterization a Problem for Profile Mixture Models? Systematic Biology. 73:53–75.

Baños H., Wong T.K.F., Daneau J., Susko E., Minh B.Q., Lanfear R., Brown M.W., Eme L., Roger A.J. 2024b. GTRpmix: A Linked General Time-Reversible Model for Profile Mixture Models. Molecular Biology and Evolution. 41:msae174.

Battistuzzi F.U., Hedges S.B. 2009. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. Molecular Biology and Evolution. 26:335–343.

Betts H.C., Puttick M.N., Clark J.W., Williams T.A., Donoghue P.C.J., Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nature Ecology and Evolution. 2:1556–1562.

Capella-Gutiérrez S., Marcet-Houben M., Gabaldón T. 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. BMC Biology. 10:47.

Castelli M., Nardi T., Gammuto L., Bellinzona G., Sabaneyeva E., Potekhin A., Serra V., Petroni G., Sassera D. 2024. Host association and intracellularity evolved multiple times independently in the Rickettsiales. Nature Communications. 15.

Castelli M., Sassera D., Petroni G. 2016. Biodiversity of "non-model" Rickettsiales and their association with aquatic organisms. Rickettsiales: Biology, Molecular Biology, Epidemiology, and Vaccine Development. p. 59–91.

Demotte P., Panchaksaram M., Kumarasinghe H., Ly-Trong N., Reis M. dos, Minh B.Q. 2025. IQ2MC: A New Framework to Infer Phylogenetic Time Trees Using IQ-TREE 3 and MCMCTree with Mixture Models. .

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology. 4:699–710.

Eme L., Tamarit D., Caceres E.F., Stairs C.W., De Anda V., Schön M.E., Seitz K.W., Dombrowski N., Lewis W.H., Homa F., Saw J.H., Lombard J., Nunoura T., Li W.-J., Hua Z.-S., Chen L.-X., Banfield J.F., John E.S., Reysenbach A.-L., Stott M.B., Schramm A.,

Kjeldsen K.U., Teske A.P., Baker B.J., Ettema T.J.G. 2023. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature. 618:992–999.

Emms D.M., Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology. 20:238.

Galindo L.J., López-García P., Torruella G., Karpov S., Moreira D. 2021. Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. Nat Commun. 12:4973.

Gaston D., Susko E., Roger A.J. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. Bioinformatics. 27:2655–2663.

Goto N., Prins P., Nakao M., Bonnal R., Aerts J., Katayama T. 2010. BioRuby: Bioinformatics software for the Ruby programming language. Bioinformatics. 26:2617–2619.

Guindon S. 2019. PhyML Manual. .

Guindon S. 2020. Rates and Rocks: Strengths and Weaknesses of Molecular Dating Methods. Front. Genet. 11.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology. 59:307–321.

Ho S.Y.W., Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular Ecology. 23:5947–5965.

Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary Biology. 7:S4.

Lartillot N., Philippe H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution. 21:1095–1109.

Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Molecular Biology and Evolution. 29:2921–2936.

Le S.Q., Gascuel O. 2010. Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. Systematic Biology. 59:277–287.

Le S.Q., Lartillot N., Gascuel O. 2008. Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society B: Biological Sciences. 363:3965–3976.

Ly-Trong N., Naser-Khdour S., Lanfear R., Minh B.Q. 2022. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era. Molecular Biology and Evolution. 39.

Mahendrarajah T.A., Moody E.R.R., Schrempf D., Szánthó L.L., Dombrowski N., Davín A.A., Pisani D., Donoghue P.C.J., Szöllősi G.J., Williams T.A., Spang A. 2023. ATP synthase evolution on a cross-braced dated tree of life. Nat Commun. 14:7456.

Marin J., Battistuzzi F.U., Brown A.C., Hedges S.B. 2017. The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation. Molecular biology and evolution. 34:437–446.

Martijn J., Vosseberg J., Guy L., Offre P., Ettema T.J.G. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. 557:101–105.

Moody E.R.R., Álvarez-Carretero S., Mahendrarajah T.A., Clark J.W., Betts H.C., Dombrowski N., Szánthó L.L., Boyle R.A., Daines S., Chen X., Lane N., Yang Z., Shields G.A., Szöllősi G.J., Spang A., Pisani D., Williams T.A., Lenton T.M., Donoghue P.C.J. 2024. The nature of the last universal common ancestor and its impact on the early Earth system. Nat Ecol Evol. 8:1654–1666.

Moody E.R.R., Mahendrarajah T.A., Dombrowski N., Clark J.W., Petitjean C., Offre P., Szöllősi G.J., Spang A., Williams T.A. 2022. An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. eLife. 11.

Pagel M., Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology. 53:571–581.

Parfrey L.W., Lahr D.J.G., Knoll A.H., Katz L.A. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences of the United States of America. 108:13624–13629.

Quang L.S., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.

Rannala B., Yang Z. 2007. Inferring speciation tunes under an episodic molecular clock. Systematic Biology. 56:453–466.

dos Reis M., Donoghue P.C.J., Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nature Reviews Genetics. 17:71–80.

dos Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian Estimation of Divergence Times. Molecular Biology and Evolution. 28:2161–2172.

Scholl J.P., Wiens J.J. 2016. Diversification rates and species richness across the Tree of Life. Proceedings of the Royal Society B: Biological Sciences. 283:20161335.

Schrempf D., Lartillot N., Szöllosi G. 2020. Scalable empirical mixture models that account for across-site compositional heterogeneity. Molecular Biology and Evolution. 37:3616–3631.

Seo T.K., Kishino H., Thorne J.L. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. Molecular Biology and Evolution. 21:1201–1213.

Stadler T. 2011. Simulating trees with a fixed number of extant species. Systematic Biology. 60:676–684.

Szánthó L.L., Lartillot N., Szöllősi G.J., Schrempf D. 2023. Compositionally Constrained Sites Drive Long-Branch Attraction. Systematic Biology. 72:767–780.

Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution. 15:1647–1657.

Venditti C., Meade A., Pagel M. 2008. Phylogenetic Mixture Models Can Reduce Node-Density Artifacts. Systematic Biology. 57:286–293.

Wang H.C., Li K., Susko E., Roger A.J. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evolutionary Biology. 8.

Wang H.C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Systematic Biology. 67:216–235.

Wang H.C., Susko E., Roger A.J. 2019. The Relative Importance of Modeling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference. Systematic Biology. 68:1003–1019.

Wang S., Luo H. 2021. Dating Alphaproteobacteria evolution with eukaryotic fossils. Nature Communications. 12:3324.

Wang S., Luo H. 2025. Dating the bacterial tree of life based on ancient symbiosis. Systematic Biology.:syae071.

Wang Z., Wu M. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Scientific Reports. 5:7949.

Williams T.A., Cox C.J., Foster P.G., Szöllősi G.J., Embley T.M. 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol. 4:138–147.

Wong T.K.F., Ly-Trong N., Ren H., Baños H., Roger A.J., Susko E., Bielow C., Maio N.D., Goldman N., Hahn M.W., Huttley G., Lanfear R., Minh B.Q. 2025. IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. .

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396–1401.

794 Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable
795 rates over sites: Approximate methods. Journal of Molecular Evolution. 39:306–314.

796 Yang Z., Rannala B. 2006. Bayesian estimation of species divergence times under a molecular
797 clock using multiple fossil calibrations with soft bounds. Molecular Biology and
798 Evolution. 23:212–226.

799 Zuckerkandl E., Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins.
800 Evolving Genes and Proteins. p. 97–166.

801

802

803

804 **Figure legends**

805 **Figure 1.** Comparison of the accuracy of divergence time (A) and branch-specific substitution
806 rate (B) estimation under substitution models LG+G and LG+C60+G by simulation. Each plot
807 contains 30 dots representing the mean relative difference based on MCMCtree molecular clock
808 dating with the corresponding calibration strategy on 30 simulated datasets. For each simulation,
809 30 timetrees each containing 20 tips are simulated using a birth-death process with a mean
810 substitution rate 0.25 substitutions/site/Gyr across four different root ages ranging from 1.0 to 4.0
811 Ga. Sequence alignments are generated with LG+C60+G4{1.0} (+G4{1.0} indicates $\alpha = 1$ in a
812 four-category discrete Gamma-distributed across-site relative rates) under AR (left-hand side)
813 and IR (right-hand side) clock models respectively. The index relative difference (*y*-axis) is used
814 to measure the accuracy of time or rate estimates as compared with the true values (see
815 Methods). *P*-values are obtained from a paired Wilcoxon signed-rank test. The box plots
816 (median, interquartile range, 1.5×IQR whiskers) illustrate the relative differences in time (A) and
817 rate (B) estimates compared to true values. The calibration schemes are given as follows. *root*: a
818 single bounded calibration (both upper and lower bounds); *calib-1_0.2--only_min*: bounded
819 calibration at the root and one minimum-bounded internal node; *calib-1_0.2*: bounded calibration
820 at both the root and an internal node; *calib-2_0.2*: bounded calibration at the root and two
821 minimum-bounded internal nodes; *calib-2_0.2--only_min*: bounded calibration at both the root
822 and two internal nodes.

823

824 **Figure 2.** Investigating the robustness of divergence time estimation to alternative settings in
825 simulation. (A) Sequence simulated under LG+C60+G4{0.5} (*Gamma_0.5*; the parameter alpha
826 in the discrete Gamma distribution equal to 1.0), thus more rate heterogeneity across sites than

827 the focal analysis (LG+C60+G4{1.0}). (B) Tree timetree simulated under a flat time prior
828 (*bd_flat*). (C) Timetree simulated under a molecular clock with larger across-branch rate variance
829 (*sigma_0.38*). The plots on the left and right correspond to analyses using the AR and IR clock
830 models, respectively. More detailed information about the above alternative simulation schemes
831 can be found in Note S1. Calibration strategies and other settings are the same as those used in
832 Fig. 1.
833
834 **Figure 3**. Evaluation of the impacts of calibration and model misspecification on molecular
835 clock dating. Accuracy of divergence time estimation (relative difference) is shown as the *y*-axis.
836 (A, B) The lower and upper time bounds of all calibrations are shifted forward (A) and backward
837 (B) in time by 20% of their true values, respectively. (C) Assessment of the accuracy of
838 divergence time estimation under alternative substitution models used in simulation (LG+C60+G
839 used in most other analyses): LG+G, LG+G+I, LG+R, LG+C20+G, LG+C60+G+I. Boxplots
840 compare the accuracy obtained by the true model (red) and by LG+C60+G (light green). Each
841 plot contains 30 dots representing the mean relative difference based on MCMCtree molecular
842 clock dating with the corresponding calibration strategy on 30 simulated datasets. For each
843 simulation, 30 timetrees each containing 20 tips are simulated using a birth-death process with a
844 mean substitution rate 0.25 substitutions/site/Gyr across four different root ages ranging from 1.0
845 to 4.0 Ga. The plots on the left and right correspond to analyses using the AR and IR clock
846 models, respectively. Calibration strategies are the same as those used in Fig. 1.
847
848 **Figure 4**. Evaluation of the approximation of PMSF and bootstrapped Hessian on molecular
849 clock analysis by simulation. Phylogenies are simulated under the same strategy as the focal

850 dataset. The sequence alignment is simulated under LG+C60+G{1.0}. The bootstrapped Hessian
851 matrix is calculated using the software bs_inBV. The *y*-axis indicates the relative accuracy of the
852 posterior mean of the divergence times (A) or substitution rates (B) compared to that obtained by
853 the first MCMC run under LG+C60+G. Each plot contains 30 dots representing the mean
854 relative difference based on MCMCtree molecular clock dating with the corresponding
855 calibration strategy on 30 simulated datasets. For each simulation, 30 timetrees each containing
856 20 tips are simulated using a birth-death process with a mean substitution rate 0.25
857 substitutions/site/Gyr across four different root ages ranging from 1.0 to 4.0 Ga.
858 *LG+C60+G_run2*: the second independent MCMC run of MCMCtree under LG+C60+G.
859 *LG+C60+G+PMSF*: PMSF approximation with LG+C60+G. *LG+C60+G+bs_inBV*:
860 bootstrapped Hessian under LG+C60+G. *LG+C60+G+PMSF+bs_inBV:* bootstrapped Hessian
861 under *LG+C60+G+PMSF*. Calibration strategies are the same as those used in Fig. 1.
862
863 **Figure 5.** Evolutionary timeline and substitution rates estimated by MCMCtree under different
864 substitution models for Microsporidia. A uniform calibration prior between 0.5 and 0.9 Ga was
865 applied to the root (see Note S2), with soft bounds extending 2.5% beyond each time bound.
866 Branch colors indicate the logarithm of the posterior mean substitution rate (estimated by
867 MCMCtree), ranging from blue (low) to red (high). (B) Posterior divergence times and branch-
868 specific substitution rates estimated by the mixture substitution model (*y*-axis) versus those
869 estimated by the homogeneous substitution model (*x*-axis) under the AR or IR clock models. The
870 bars indicate the 95% HPD.
871

872 **Figure 6.** Evolutionary timeline and substitution rates estimated by MCMCtree under different
873 substitution models for Rickettsiales. This is based the mitochondrial endosymbiosis dating
874 strategy (Wang and Luo 2021), using eukaryotic fossils to calibrate Rickettsiales evolution,
875 based on the evolutionary scenario that mitochondria originated from a lineage closely related to
876 α-Proteobacteria. (A) The estimated evolutionary timeline of Rickettsiales under different
877 substitution (the best-fitting homogeneous or mixture) and clock (IR or AR) models. All
878 calibrated nodes are denoted by a circle. In addition to the calibration at the root, all four
879 calibrations are placed within eukaryotes in the clade of mitochondria (see Note S2). Branch
880 colors indicate the logarithm of the posterior mean substitution rate estimated by MCMCtree,
881 ranging from blue (low) to red (high). (B) Posterior divergence times and branch-specific
882 substitution rates estimated by the mixture substitution model (*y*-axis) versus those estimated by
883 the homogeneous substitution model (*x*-axis) under the AR or IR clock models. The bars indicate
884 the 95% HPD.
885
886

Figure 1

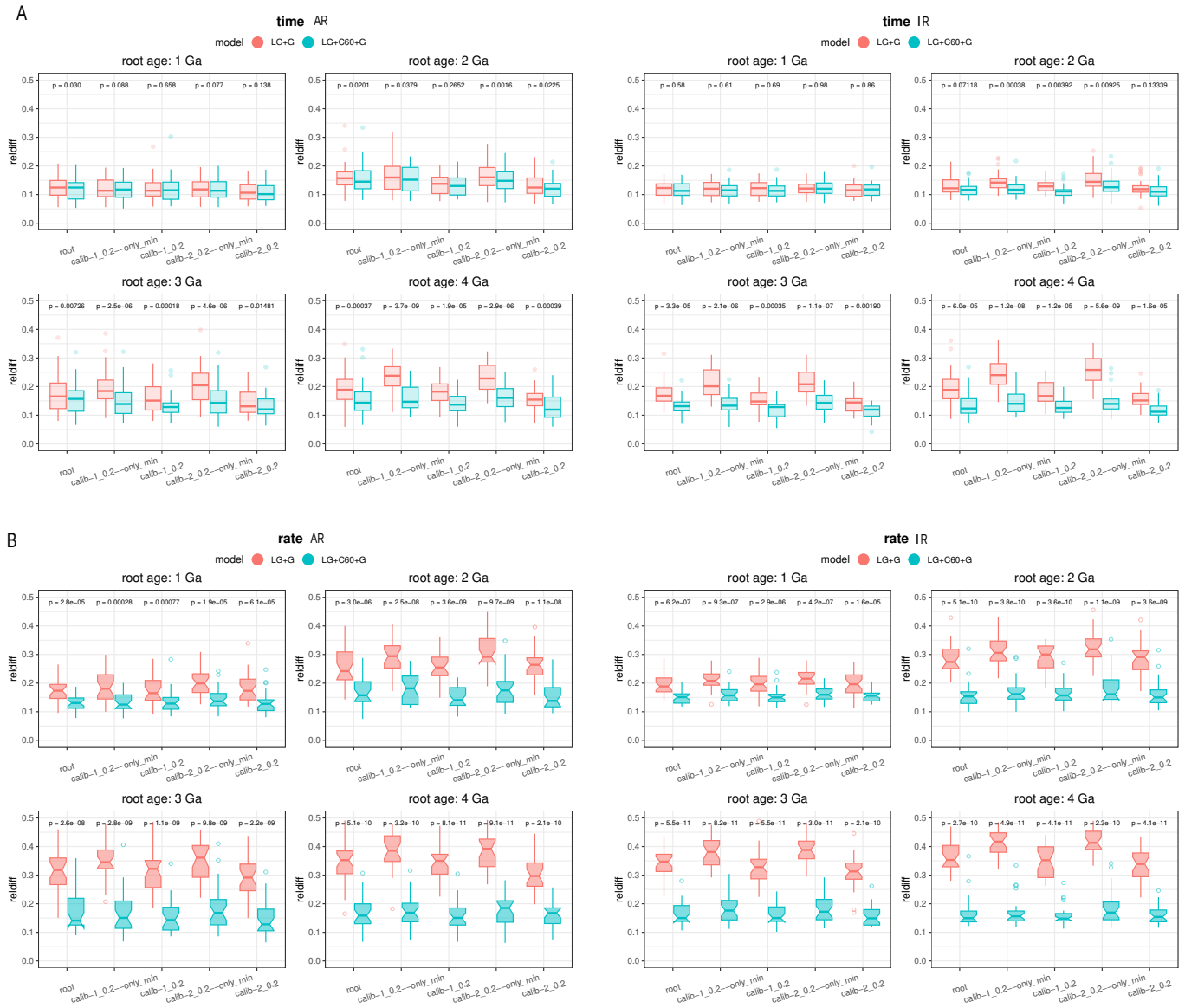

A
time AR
model LG+G LG+C60+G
root age: 1 Ga
p = 0.030 p = 0.088 p = 0.658 p = 0.077 p = 0.138
root age: 2 Ga
p = 0.0201 p = 0.0379 p = 0.2652 p = 0.0016 p = 0.0225
root age: 3 Ga
p = 0.00726 p = 2.5e−06 p = 0.00018 p = 4.6e−06 p = 0.01481
root age: 4 Ga
p = 0.00037 p = 3.7e−09 p = 1.9e−05 p = 2.9e−06 p = 0.00039
time IR
root age: 1 Ga
p = 0.58 p = 0.61 p = 0.69 p = 0.98 p = 0.86
root age: 2 Ga
p = 0.07118 p = 0.00038 p = 0.00392 p = 0.00925 p = 0.13339
root age: 3 Ga
p = 3.3e−05 p = 2.1e−06 p = 0.00035 p = 1.1e−07 p = 0.00190
root age: 4 Ga
p = 6.0e−05 p = 1.2e−08 p = 1.2e−05 p = 5.6e−09 p = 1.6e−05
reldiff
root calib−1_0.2 calib−1_0.2−only_min calib−2_0.2 calib−2_0.2−only_min calib−2_0.2
B
rate AR
root age: 1 Ga
p = 2.8e−05 p = 0.00028 p = 0.00077 p = 1.9e−05 p = 6.1e−05
root age: 2 Ga
p = 3.0e−06 p = 2.5e−08 p = 3.6e−09 p = 9.7e−09 p = 1.1e−08
root age: 3 Ga
p = 2.6e−08 p = 2.8e−09 p = 1.1e−09 p = 9.8e−09 p = 2.2e−09
root age: 4 Ga
p = 5.1e−10 p = 3.2e−10 p = 8.1e−11 p = 9.1e−11 p = 2.1e−10
rate IR
root age: 1 Ga
p = 6.2e−07 p = 9.3e−07 p = 2.9e−06 p = 4.2e−07 p = 1.6e−05
root age: 2 Ga
p = 5.1e−10 p = 3.8e−10 p = 3.6e−10 p = 1.1e−09 p = 3.6e−09
root age: 3 Ga
p = 5.5e−11 p = 8.2e−11 p = 5.5e−11 p = 3.0e−11 p = 2.1e−10
root age: 4 Ga
p = 2.7e−10 p = 4.9e−11 p = 4.1e−11 p = 2.3e−10 p = 4.1e−11

Figure 2

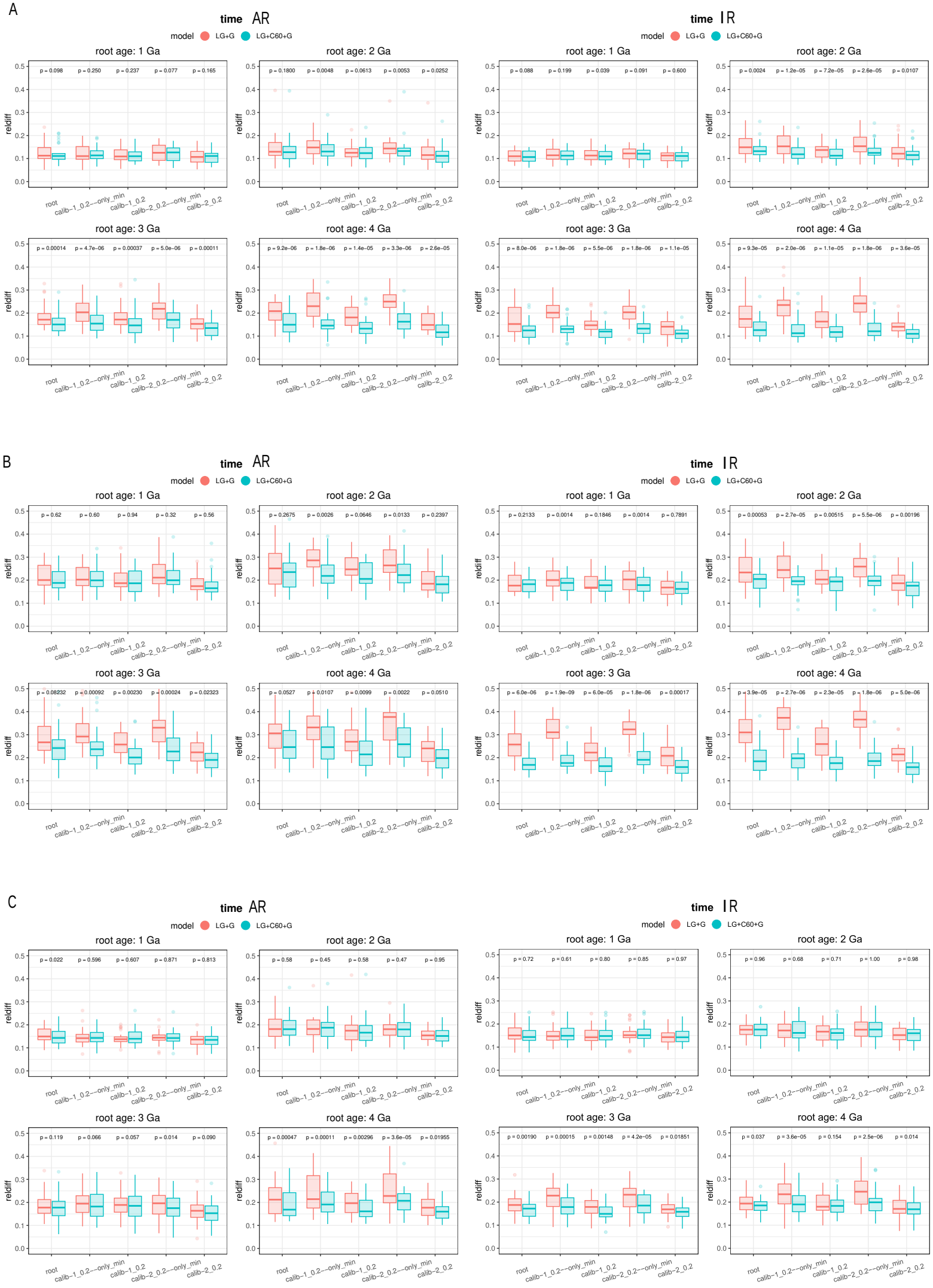

A
time AR
model LG+G LG+C60+G
root age: 1 Ga
p = 0.098 p = 0.250 p = 0.237 p = 0.077 p = 0.165
root age: 2 Ga
p = 0.1800 p = 0.0048 p = 0.0613 p = 0.0053 p = 0.0252
root age: 3 Ga
p = 0.00014 p = 4.7e−06 p = 0.00037 p = 5.0e−06 p = 0.00011
root age: 4 Ga
p = 9.2e−06 p = 1.8e−06 p = 1.4e−05 p = 3.3e−06 p = 2.6e−05
time IR
root age: 1 Ga
p = 0.088 p = 0.199 p = 0.039 p = 0.091 p = 0.600
root age: 2 Ga
p = 0.0024 p = 1.2e−05 p = 7.2e−05 p = 2.6e−05 p = 0.0107
root age: 3 Ga
p = 8.0e−06 p = 1.8e−06 p = 5.5e−06 p = 1.8e−06 p = 1.1e−05
root age: 4 Ga
p = 9.3e−05 p = 2.0e−06 p = 1.1e−05 p = 1.8e−06 p = 3.6e−05
B
time AR
root age: 1 Ga
p = 0.62 p = 0.60 p = 0.94 p = 0.32 p = 0.56
root age: 2 Ga
p = 0.2675 p = 0.0026 p = 0.0646 p = 0.0133 p = 0.2397
root age: 3 Ga
p = 0.08232 p = 0.00092 p = 0.00230 p = 0.00024 p = 0.02323
root age: 4 Ga
p = 0.0527 p = 0.0107 p = 0.0099 p = 0.0022 p = 0.0510
time IR
root age: 1 Ga
p = 0.2133 p = 0.0014 p = 0.1846 p = 0.0014 p = 0.7891
root age: 2 Ga
p = 0.00053 p = 2.7e−05 p = 0.00515 p = 5.5e−06 p = 0.00196
root age: 3 Ga
p = 6.0e−06 p = 1.9e−09 p = 6.0e−05 p = 1.8e−06 p = 0.00017
root age: 4 Ga
p = 3.9e−05 p = 2.7e−06 p = 2.3e−05 p = 1.8e−06 p = 5.0e−06
C
time AR
root age: 1 Ga
p = 0.022 p = 0.596 p = 0.607 p = 0.871 p = 0.813
root age: 2 Ga
p = 0.58 p = 0.45 p = 0.58 p = 0.47 p = 0.95
root age: 3 Ga
p = 0.119 p = 0.066 p = 0.057 p = 0.014 p = 0.090
root age: 4 Ga
p = 0.00047 p = 0.00011 p = 0.00296 p = 3.6e−05 p = 0.01955
time IR
root age: 1 Ga
p = 0.72 p = 0.61 p = 0.80 p = 0.85 p = 0.97
root age: 2 Ga
p = 0.96 p = 0.68 p = 0.71 p = 1.00 p = 0.98
root age: 3 Ga
p = 0.00190 p = 0.00015 p = 0.00148 p = 4.2e−05 p = 0.01851
root age: 4 Ga
p = 0.037 p = 3.6e−05 p = 0.154 p = 2.5e−06 p = 0.014
reldiff
root
calib−1_0.2−−only_min
calib−1_0.2
calib−2_0.2−−only_min
calib−2_0.2

Figure 3

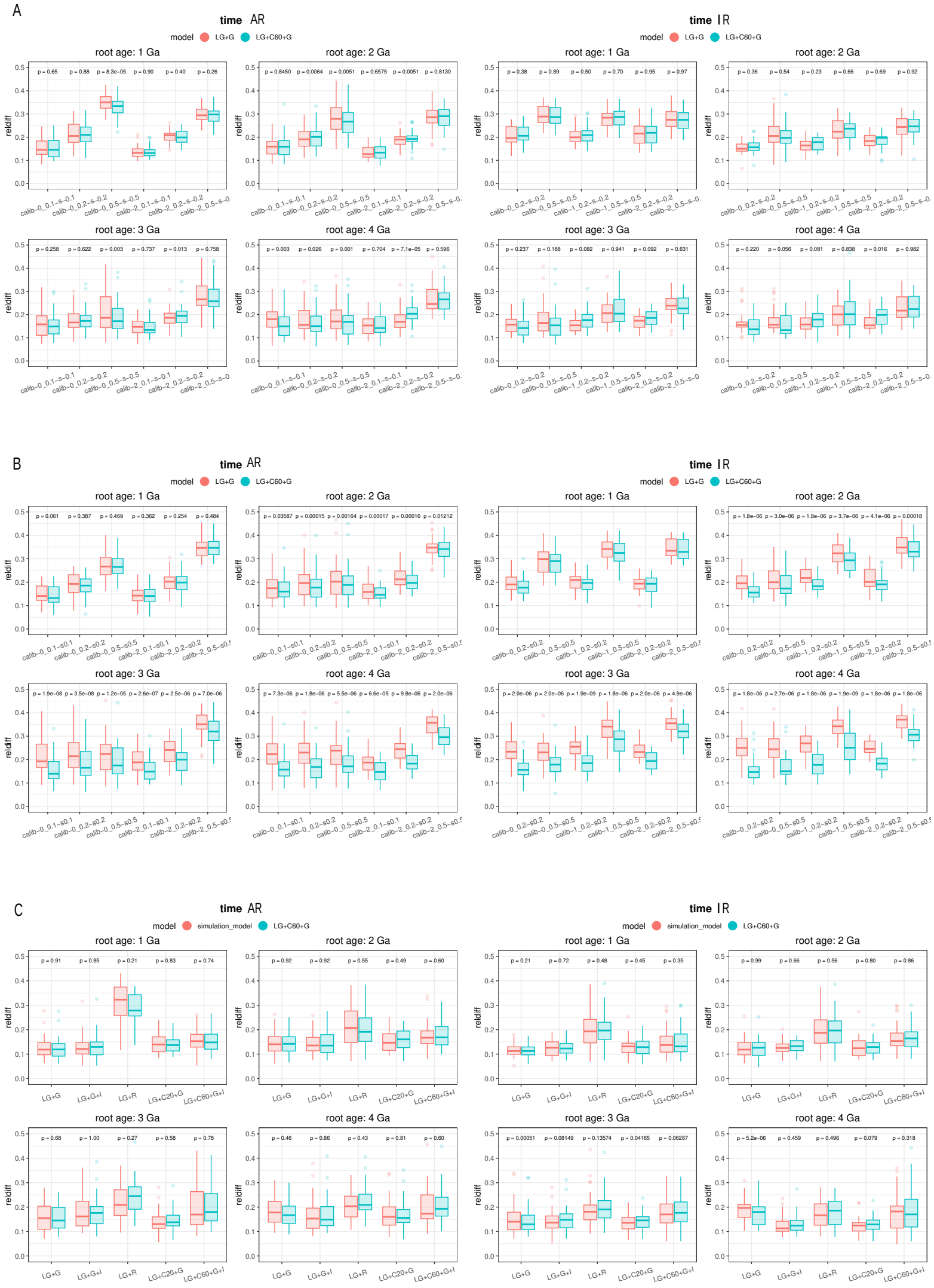

A
time AR
model LG+G LG+C60+G
root age: 1 Ga
p = 0.65 p = 0.88 p = 8.3e−05 p = 0.90 p = 0.40 p = 0.26
root age: 2 Ga
p = 0.8450 p = 0.0064 p = 0.0051 p = 0.6575 p = 0.0051 p = 0.8130
root age: 3 Ga
p = 0.258 p = 0.622 p = 0.003 p = 0.737 p = 0.013 p = 0.758
root age: 4 Ga
p = 0.003 p = 0.026 p = 0.001 p = 0.704 p = 7.1e−05 p = 0.596
time IR
root age: 1 Ga
p = 0.38 p = 0.89 p = 0.50 p = 0.70 p = 0.95 p = 0.97
root age: 2 Ga
p = 0.36 p = 0.54 p = 0.23 p = 0.66 p = 0.69 p = 0.92
root age: 3 Ga
p = 0.237 p = 0.188 p = 0.082 p = 0.941 p = 0.092 p = 0.631
root age: 4 Ga
p = 0.220 p = 0.056 p = 0.081 p = 0.838 p = 0.016 p = 0.982
B
time AR
root age: 1 Ga
p = 0.061 p = 0.387 p = 0.469 p = 0.362 p = 0.254 p = 0.484
root age: 2 Ga
p = 0.03587 p = 0.00015 p = 0.00164 p = 0.00017 p = 0.00016 p = 0.01212
root age: 3 Ga
p = 1.9e−08 p = 3.5e−08 p = 1.2e−05 p = 2.6e−07 p = 2.5e−06 p = 7.0e−06
root age: 4 Ga
p = 7.3e−06 p = 1.8e−06 p = 5.5e−06 p = 6.6e−05 p = 9.8e−06 p = 2.0e−06
time IR
root age: 1 Ga
root age: 2 Ga
p = 1.8e−06 p = 3.0e−06 p = 1.8e−06 p = 3.7e−06 p = 4.1e−06 p = 0.00018
root age: 3 Ga
p = 2.0e−06 p = 2.0e−06 p = 1.9e−09 p = 1.8e−06 p = 2.0e−06 p = 4.9e−06
root age: 4 Ga
p = 1.8e−06 p = 2.7e−06 p = 1.8e−06 p = 1.9e−09 p = 1.8e−06 p = 1.8e−06
C
time AR
model simulation_model LG+C60+G
root age: 1 Ga
p = 0.91 p = 0.85 p = 0.21 p = 0.83 p = 0.74
root age: 2 Ga
p = 0.92 p = 0.92 p = 0.55 p = 0.49 p = 0.60
root age: 3 Ga
p = 0.68 p = 1.00 p = 0.27 p = 0.58 p = 0.78
root age: 4 Ga
p = 0.46 p = 0.86 p = 0.43 p = 0.81 p = 0.60
time IR
root age: 1 Ga
p = 0.21 p = 0.72 p = 0.48 p = 0.45 p = 0.35
root age: 2 Ga
p = 0.99 p = 0.66 p = 0.56 p = 0.80 p = 0.86
root age: 3 Ga
p = 0.00051 p = 0.08149 p = 0.13574 p = 0.04165 p = 0.06287
root age: 4 Ga
p = 5.2e−06 p = 0.459 p = 0.496 p = 0.079 p = 0.318
reldiff
LG+G LG+G+I LG+R LG+C20+G LG+C60+G+I

Figure 4

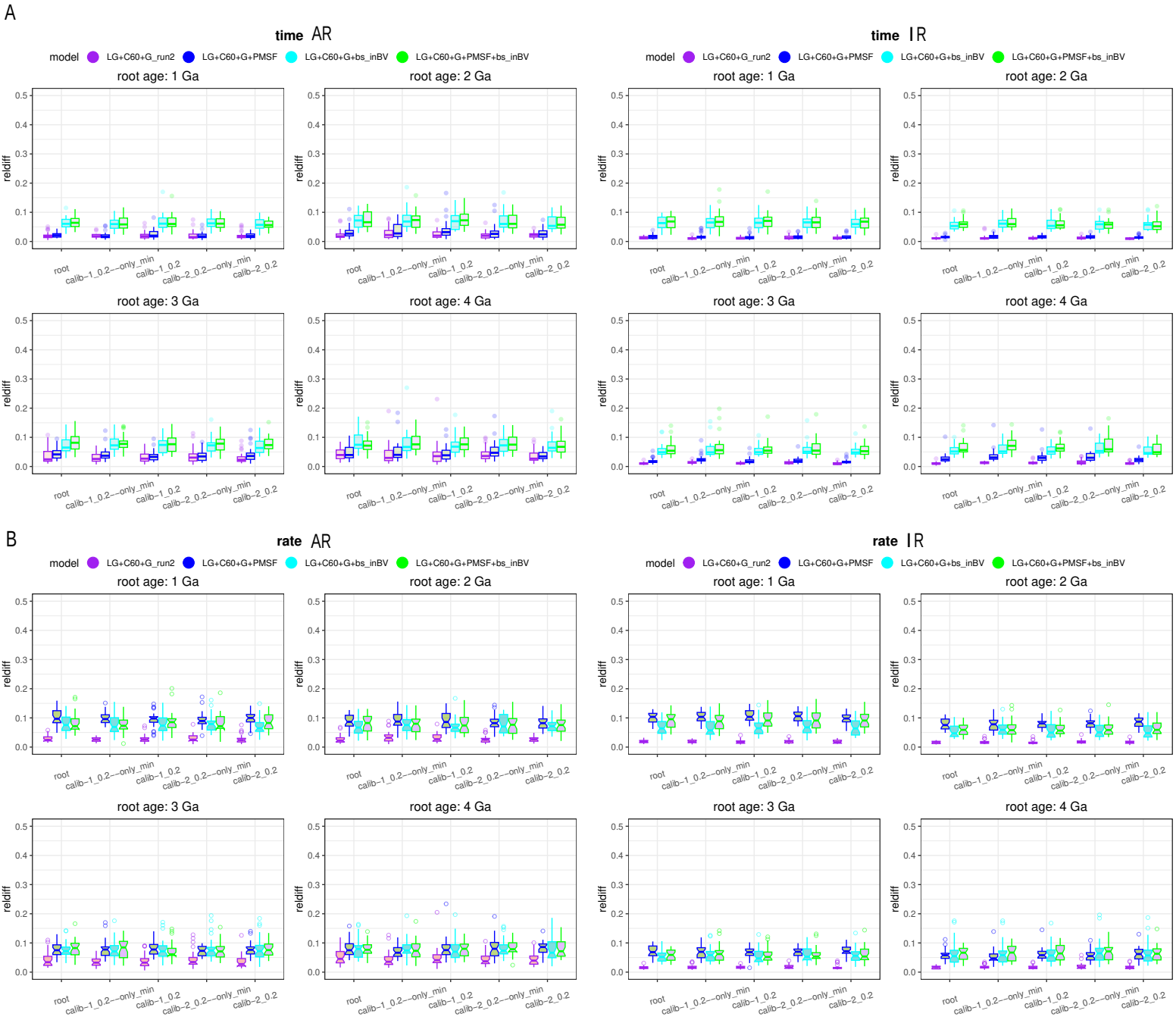
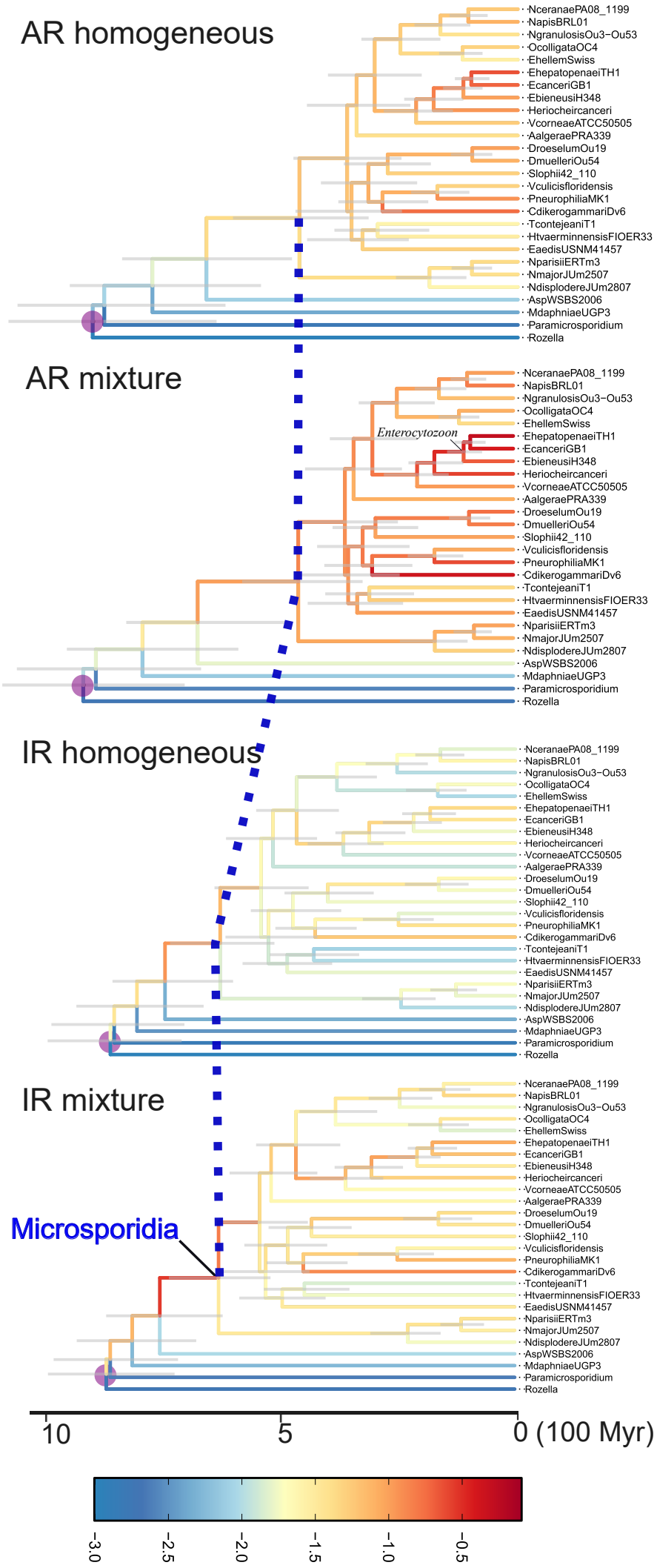

A
time AR
time IR
B
rate AR
rate IR
model
LG+C60+G_run2
LG+C60+G+PMSF
LG+C60+G+bs_inBV
LG+C60+G+PMSF+bs_inBV
root age: 1 Ga
root age: 2 Ga
root age: 3 Ga
root age: 4 Ga
reldiff
0.0
0.1
0.2
0.3
0.4
0.5
root
calib−1_0.2−−only_min
calib−1_0.2
calib−2_0.2−−only_min
calib−2_0.2

Figure 5

A


AR homogeneous
AR mixture
IR homogeneous
IR mixture
Enterocytozoon
Microsporidia
NceranaePA08_1199
NapisBRL01
NgranulosisOu3−Ou53
OcolligataOC4
EhellemSwiss
EhepatopenaeiTH1
EcanceriGB1
EbieneusiH348
Heriocheircanceri
VcorneaeATCC50505
AalgeraePRA339
DroeselumOu19
DmuelleriOu54
Slophii42_110
Vculicisfloridensis
PneurophiliaMK1
CdikerogammariDv6
TcontejeaniT1
HtvaerminnensisFIOER33
EaedisUSNM41457
NparisiiERTm3
NmajorJUm2507
NdisplodereJUm2807
AspWSBS2006
MdaphniaeUGP3
Paramicrosporidium
Rozella
10
5
0 (100 Myr)
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5

B

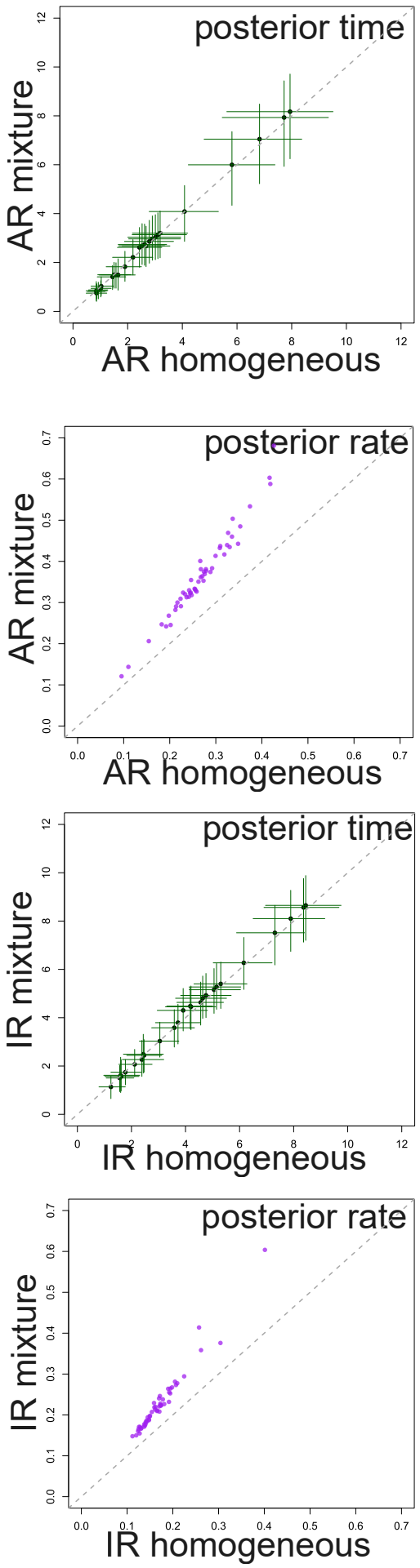
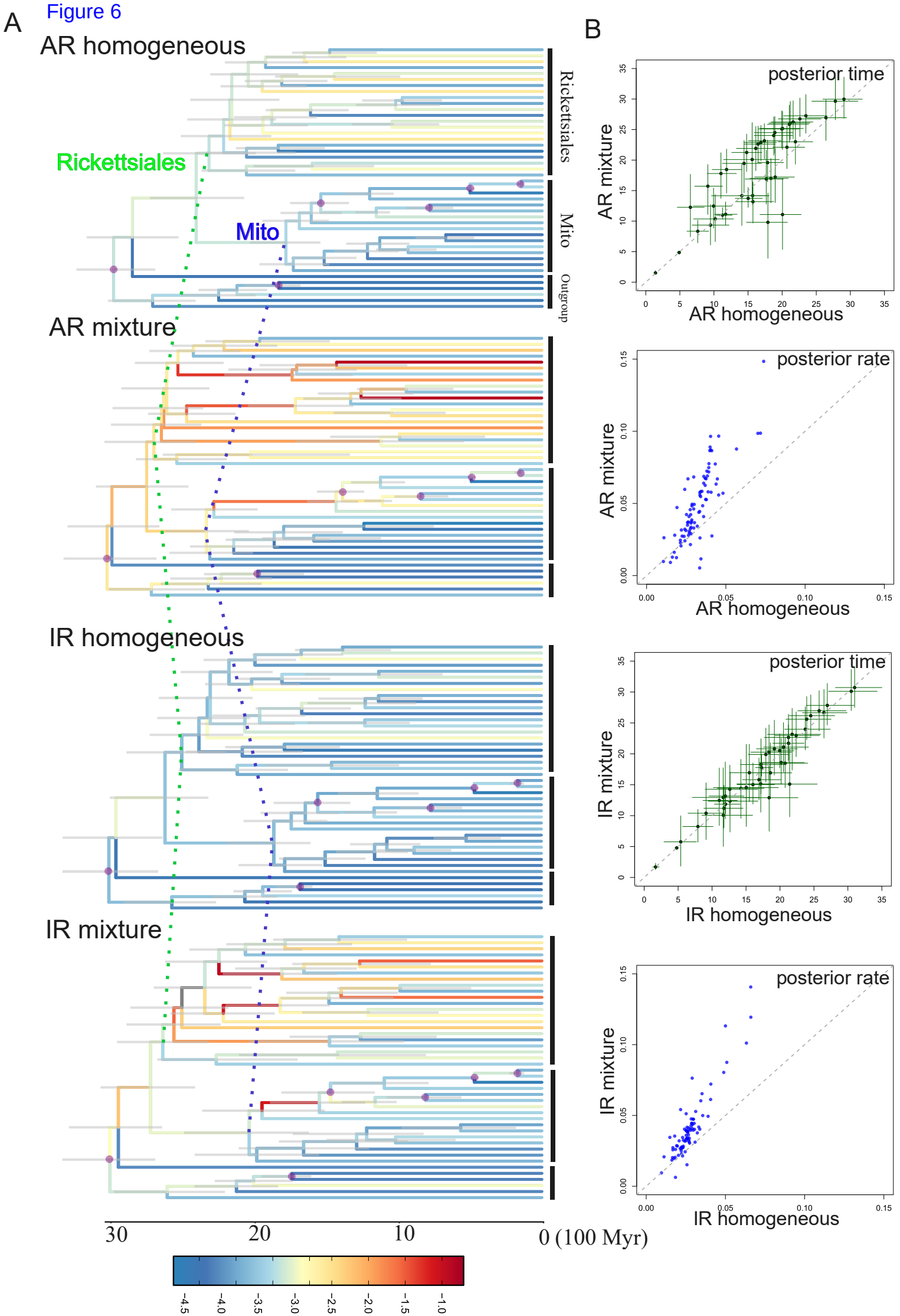

posterior time
AR mixture
AR homogeneous
posterior rate
AR mixture
AR homogeneous
posterior time
IR mixture
IR homogeneous
posterior rate
IR mixture
IR homogeneous

Figure 6


A
AR homogeneous
Rickettsiales
Mito
Rickettsiales
Mito
Outgroup
AR mixture
IR homogeneous
IR mixture
30
20
10
0 (100 Myr)
-4.5
-4.0
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
B
posterior time
AR mixture
AR homogeneous
posterior rate
AR mixture
AR homogeneous
posterior time
IR mixture
IR homogeneous
posterior rate
IR mixture
IR homogeneous