

Dating the Bacterial Tree of Life Based on Ancient Symbiosis

SISHUO WANG^{1,2} , AND HAIWEI LUO^{1,3,*} 

¹Simon F. S. Li Marine Science Laboratory, School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Tai Po Road, Shatin, Hong Kong SAR

²Department of Microbiology, The Chinese University of Hong Kong, Tai Po Road, Shatin, Hong Kong SAR

³Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Tai Po Road, Shatin, Hong Kong SAR

*Correspondence to be sent to: School of Life Sciences, The Chinese University of Hong Kong, Tai Po Road, Hong Kong SAR; E-mail: haiweiluo@cuhk.edu.hk.

Received 27 November 2023; reviews returned 31 October 2024; accepted 29 November 2024

Associate Editor: Ziheng Yang

Abstract.—Obtaining a timescale for bacterial evolution is crucial to understand early life evolution but is difficult owing to the scarcity of bacterial fossils. Here, we introduce multiple new time constraints to calibrate bacterial evolution based on ancient symbiosis. This idea is implemented using a bacterial tree constructed with genes found in the mitochondrial lineages phylogenetically embedded within Proteobacteria. The expanded mitochondria-bacterial tree allows the node age constraints of eukaryotes established by their abundant fossils to be propagated to ancient co-evolving bacterial symbionts and across the bacterial tree of life. Importantly, we formulate a new probabilistic framework that considers uncertainty in inference of the ancestral lifestyle of modern symbionts to apply 19 relative time constraints each informed by host-symbiont association to constrain bacterial symbionts no older than their eukaryotic host. Moreover, we develop an approach to incorporating substitution mixture models that better accommodate substitutional saturation and compositional heterogeneity for dating deep phylogenies. Our analysis estimates that the last bacterial common ancestor occurred approximately 4.0–3.5 billion years ago (Ga), followed by rapid divergence of major bacterial clades. It is generally robust to alternative root ages, root positions, tree topologies, fossil ages, ancestral lifestyle reconstruction, gene sets, among other factors. The obtained timetree serves as a foundation for testing hypotheses regarding bacterial diversification and its correlation with geobiological events across different timescales. [keyword: Molecular clock; bacterial tree of life; ancestral reconstruction of lifestyles; timetree; symbiosis; relative time constraint.]

Bacteria are the dominating form of life in the first two-thirds of the history of our planet, and their evolution and diversification have greatly shaped the geosphere of Earth (Falkowski et al. 2008). Many hypotheses have been proposed to interpret outstanding questions like how bacteria affected the evolution of biogeochemical cycles or when major innovations of bacteria such as photosynthesis occurred (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009; Louca et al. 2018). Interpreting these questions depends on temporal links between bacterial evolution and the corresponding geobiological events. However, efforts to date the early evolution of bacteria based on relaxed molecular clocks (Zuckerandl and Pauling 1965; Drummond et al. 2006; Yang and Rannala 2006) often come to different conclusions. A major challenge is that biomarkers or fossils attributable to bacteria are scarce, and thus the nodes whose age can be constrained for bacterial timetree inference are few (Battistuzzi and Hedges 2009). In addition, because fossils or biomarkers by themselves provide only a lower minimum bound (Benton and Donoghue 2007; Marshall 2019), there are no reliable maximum time constraints to calibrate bacterial evolution. Accordingly, the time estimates heavily depend on the maximum age assigned to the root, that is, the last bacterial common ancestor (LBCA), whose origin time involves even more debate (Wang and Luo 2021; Moody et al. 2022). Moreover, previous studies in dating the early evolution of bacteria (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009; Wang and Luo 2021) are mostly based

on site-homogeneous amino acid substitution models that do not account for among-site compositional heterogeneity or substitutional saturation (Gouy et al. 2015). In deep-time phylogenetics, this may lead to an underestimate of branch length (expected number of substitutions per site) (Moody et al. 2022), which potentially affects time estimation.

Recent studies have used eukaryotic fossils to date the evolution of Proteobacteria (Wang and Luo 2021), ammonia-oxidizing bacteria and anaerobic ammonium-oxidizing bacteria (Liao et al. 2022, 2024), and Cyanobacteria (Sánchez-Baracaldo et al. 2017; Shih et al. 2017; Zhang et al. 2023) where mitochondrial (mito-) and/or plastid-endosymbiosis are involved. These successes imply the value of the abundant eukaryotic fossils in timing bacterial evolution using molecular clock-dating approaches. Here, we exploit a novel type of age constraint derived from bacterial endosymbiosis, which often bears a signature of a co-evolutionary history between symbiotic bacteria and their eukaryotic hosts. Unlike mitochondrial and plastid-endosymbiosis which are phylogenetically restricted to Proteobacteria and Cyanobacteria, respectively, bacterial endosymbiosis independently occurred in multiple ancient bacterial lineages (Toft and Andersson 2010; Perreau and Moran 2022), making it possible to propagate this novel timing information to the entire bacterial tree of life.

A recently developed framework called relative time constraint (RTC) dating (Davín et al. 2018; Magnabosco

et al. 2018; Szöllösi et al. 2022) seeks to improve divergence time estimation by constraining the order of divergence by assuming the age of some nodes of the tree to be younger than others. In previous works (Davín et al. 2018; Magnabosco et al. 2018; Szöllösi et al. 2022), the order of divergence is often provided by horizontal gene transfer (HGT) where the recipient must be no older than the donor. In the present study and in the context of ancient bacterial symbiosis, RTCs may be based on the assumption that bacterial symbionts are no older than their eukaryotic hosts instead. Despite some encouraging progress, the previous implementation of this method treats all RTCs as if they are known with full certainty. However, this assumption might be oversimplified because of the uncertainty in identifying the relative time order based on HGT (Ravenhall et al. 2015) or other evolutionary events.

Here, we address the above challenges by leveraging combined evidence of molecular sequences, fossils, and ancient symbiosis to calibrate an integrative timescale of bacterial evolution. We developed and implemented a framework to propagate node age constraints of eukaryotes established by abundant fossils to ancient co-evolving bacterial symbionts and other bacteria by requiring that the origin of symbionts postdates their hosts' while accommodating the uncertainty in ancestral host reconstruction. We established timelines of bacterial evolution at different taxonomy resolutions and discussed how they may affect our views of the evolution of early life and habitat.

METHODS

Phylogenomic Reconstruction

A total of 62 gene markers considered conserved across bacteria were retrieved from a recent phylogenomics study of the bacterial tree (Coleman et al. 2021). Their sequences were aligned using MAFFT v7.22261 (Katoh and Standley 2013) and trimmed using TrimAl v1.4 (settings: “-resoverlap 0.55 -seqoverlap 60”) (Capella-Gutiérrez et al. 2009). Taxonomy sampling is detailed in Supplementary Table S1 and Supplementary Note S1 (all Supplementary Material can be found in the Dryad data repository: <https://doi.org/10.5061/dryad.1c59zw42s>). Unrooted phylogenomic trees were constructed based on the concatenated alignment with time-reversible Markov model and each of the 3 empirical site-heterogeneous profile mixture models C20, C40, and C60 using a PMSF (posterior mean site frequency) approximation (guide tree built by LG4M + G + I) with 1000 ultrafast bootstraps (Minh et al. 2020). The model LG + G + C60 was selected as the best-fit model according to AIC (Akaike information criterion) and BIC (Bayesian information criterion), and the tree constructed by this model was used in the main analysis. Rooted bacterial trees were constructed using the recently developed time-nonreversible model implemented in IQ-Tree v2.1.3 (Naser-Khdour et al. 2022) that took the unrooted tree as either a fixed tree or a starting

tree (see Supplementary Table S2 and Supplementary Note S2 for details).

Identification of Mitochondrial Genes Conserved Across the Bacterial Tree

We retrieved the sequences of the 108 genes identified to have a mitochondrial origin from the study carried out by Muñoz-Gómez et al. (2022). We identified their homologs in eukaryotes encoded by either the mitochondrial or nuclear genome based on BLAST v2.6.0 + with an e-value cutoff of $1e-30$. Next, we followed Liao et al. (2022) to keep 32 genes present in at least one of the following 3 gene sets namely bac120 (120 genes present in nearly all bacteria that commonly serve as phylogenetic markers for distinguishing between bacterial lineages) (Parks et al. 2017), Battistuzzi2009 (25 genes) (Battistuzzi and Hedges 2009), and Coleman2021 (62 genes) (Coleman et al. 2021), which include orthologs well conserved and likely vertically transmitted across bacteria. Note that 2 of the 62 orthologs originally included in Coleman2021 correspond to K04485, and were both discarded. For each of the remaining 32 genes, we built a gene tree using IQ-Tree using the model LG + G + I, based on which we further removed 1) 2 genes that did not show monophyly between mitochondria and α -proteobacteria, indicating unresolved paralogy and 2) 11 genes where mitochondria/ α -proteobacteria were placed at a basal position to all other bacteria with different rooting approaches, indicating fast-evolving genes. This returned 19 mitochondrial (some encoded by the nuclear genome) genes conserved across the bacterial tree, which were used in the main molecular clock analysis (Supplementary Table S3 and Supplementary Fig. S1).

Ancestral Lifestyle Reconstruction

The 16S rRNA genes of each analyzed symbiont group were collected from the NCBI Entrez Nucleotide database with the keywords “taxon_name[ORGN] AND 16S[TITL] AND rrna AND 1000:2000[SLEN] NOT spacer NOT trna” (April 2022). We also retrieved 16S rRNA genes predicted by RNAmmer v1.2 (Lagesen et al. 2007) from genomes deposited in Genbank (August 2021). Those without any information of host or isolation source in the metadata were removed (see Supplementary Materials). All remaining sequences were clustered using CD-HIT v4.7 (Li and Godzik 2006) with 98.7% as the identity cutoff (Yarza et al. 2014) into operational taxonomic units (OTUs), and 1 representative was chosen from each OTU (see Supplementary Materials for alternative strategies in sequence selection). Phylogenies were then reconstructed using IQ-Tree v2.1.3 with the model GTR + G + I based on 16S rRNA alignment performed by MAFFT v7.22261. We applied the stochastic character mapping (SCM) (Huelsenbeck et al. 2003) to directly sample from the joint posterior distribution, using *make.simmap* from the R package “phytools” v0.2-40 (Revell 2024) to estimate ancestral hosts (or lifestyles) of modern symbionts. To account

for phylogenetic uncertainty, we applied the so-called most-recent-common-ancestor approach (Pagel et al. 2004) by identifying the last common ancestor (LCA) of the group of species of interest, reconstructing the state of that node, and then summarizing the information across 100 ultrafast bootstrap (UFB) trees reconstructed by IQ-Tree (settings: -m GTR + G). For each of the 100 UFB trees, 100 simulated stochastic character maps were generated using the MCMC (Markov chain Monte Carlo) algorithm implemented in phyttools under the “all-rates different” model which allows transition rates at all directions to be different, totaling $100 \times 100 = 10,000$ simulations. All ancestral state reconstruction (ASR) analyses were performed on phylogram—a phylogeny where branch lengths indicate the expected number of substitutions per site, built by the 16S rRNA gene, as done in previous studies (Wang and Luo 2021).

Specifically, let S_i be the inferred lifestyle (or state) of the i th internal node of symbionts, which can take any possible hosts observed in the group of modern symbionts (including free living), denoted as s_i . By SCM, the marginal probability that the i th internal node of symbionts takes the state s_i may be approximated by

$$P(S_i = s_i) \approx \frac{n_{\{S_i=s_i\}}}{n}, \quad (1)$$

where $n_{\{S_i=s_i\}}$ is the number of times that node i takes the state s_i during n rounds of SCMs. Likewise, the same approach can be extended to estimate the joint probability that internal nodes 1 to k take states $s_1, \dots, s_k$, respectively by

$$P(S_1 = s_1, \dots, S_k = s_k) \approx \frac{n_{\{S_1=s_1, \dots, S_k=s_k\}}}{n}, \quad (2)$$

where $n_{\{S_1=s_1, \dots, S_k=s_k\}}$ is the number that the event $\{S_1 = s_1, \dots, S_k = s_k\}$ happens during all of the n rounds of SCMs. More details of the SCM-based ASR inference are given in [Supplementary Note S3.1.2](#).

pRTC (probability-based RTC) Informed by Host-Symbiont Association

Denote the data (i.e., sequence alignments, which in the present study denote the concatenated gene alignment) for molecular dating as D , and the divergence times as $t = (t_1, \dots, t_k)$ where $t_1, \dots, t_k$ indicate the divergence time of the k ancestral nodes of modern symbionts. Let S_i be the inferred lifestyle (or state) of the i th ancestral node of symbionts, which can take any possible hosts observed in the group of modern symbionts or free-living, denoted as s_i . Let A indicate the event that the origin time of all symbionts is no earlier than that of their host, that is, $A = \{t_1 \leq t_{s_1}, \dots, t_k \leq t_{s_k}\}$. It follows that

$$\begin{aligned} f(t|D, A) &\propto f(t|D) P(A|t, D) \\ &= f(t|D) P(A|t) \end{aligned} \quad (3)$$

where $f(t|D, A)$ is the posterior of divergence times t given sequence data D and the assumption A , $f(t|D)$ is the posterior of divergence times t given only sequence data D , and $P(A|t)$ indicates the probability

A is satisfied given time t . We further assume that the inference of ancestral lifestyles of modern symbionts depends on only the phylogenetic comparative data estimated by SCM and is independent of time, that is, $P(S_1 = s_1, \dots, S_k = s_k|t) = P(S_1 = s_1, \dots, S_k = s_k)$. The second term on the right-hand side of Eq. (3) can be rewritten as

$$\begin{aligned} P(A|t) &= P(t_1 \leq t_{s_1}, \dots, t_k \leq t_{s_k}|t) \\ &= \sum_{s_1, \dots, s_k} P(t_1 \leq t_{s_1}, \dots, t_k \leq t_{s_k}|t, S_1 = s_1, \dots, S_k = s_k) \\ &\quad P(S_1 = s_1, \dots, S_k = s_k|t) \\ &= \sum_{s_1, \dots, s_k} P(t_1 \leq t_{s_1}, \dots, t_k \leq t_{s_k}|t, s_1, \dots, s_k) P(S_1 = s_1, \dots, S_k = s_k) \\ &= \sum_{s_1, \dots, s_k} \left(\prod_{i=1}^k I(t_i, t_{s_i}) \right) P(S_1 = s_1, \dots, S_k = s_k), \end{aligned}$$

where $I(t_i, t_{s_i})$ is defined as

$$I(t_i, t_{s_i}) = \begin{cases} 1, & t_i \leq t_{s_i} \\ 0, & t_i > t_{s_i} \end{cases}.$$

Particularly, we define $I(t_i, t_{fl}) = 1$ (free-living ancestral state) since the RTC that symbiont's age is no older than the host's is automatically satisfied for free-living bacteria.

Furthermore, assuming that the states of the ancestral nodes of symbiotic bacteria are independent of each other so that $P(S_1 = s_1, \dots, S_k = s_k) = \prod_{i=1}^k P(S_i = s_i)$, which generally holds if the ancestral nodes are not phylogenetically too closely related (see [Supplementary Fig. S2b](#) for details), the above can be rewritten as

$$P(A|t) = \prod_{i=1}^k \sum_{s_i} I(t_i, t_{s_i}) P(S_i = s_i). \quad (4)$$

Collectively, by introducing Eq. (4) into Eq. (3), we have

$$f(t|D, A) \propto f(t|D) \left[\prod_{i=1}^k \sum_{s_i} I(t_i, t_{s_i}) P(S_i = s_i) \right]. \quad (5)$$

The above indicates that estimating the posterior time estimates under the assumption that the origin time of all symbionts is no earlier than their hosts' while considering the uncertainty in the inference of ancestral hosts can be obtained by:

- i) Perform ASR to obtain the probability of each ancestral state for selected nodes in the group of modern symbionts of interest.
- ii) Perform Bayesian molecular clock-dating analysis with fossil calibrations but without any RTC.
- iii) For each MCMC posterior sample collected from Step 2), perform a rejection sampling by accepting it with a probability according to Eq. (5) (based on the ASR results obtained from Step 1).

Obviously, Step 2 is independent of Step 1, and Step 3 is based on Step 1 and Step 2.

Simulation-Based Comparison of Time Estimates With and Without pRTC

To show that the host-bacteria association-informed pRTC framework can improve divergence time estimation, we conducted a number of simulations. We first simulated a timetree with 26 tips, 20 bacteria, 5 eukaryotes, and 1 outgroup. The ages of the root and the origin time of the host (i.e., the LCA of euk1 and euk2 in [Supplementary Figs. S3 and S4](#)) were fixed as 2.0 Ga and 1.01 Ga, respectively. The bacterial subtree (cyan clades in [Supplementary Fig. S3](#)) was simulated under a birth-death process ([Yang and Rannala 1997](#)) (birth rate = 0.4, death rate = 0.2 [lineages/100 Ma], sampling proportion = 0.1; see also [Supplementary Note S4.1](#)) with the LCA of bacteria occurring at 1.0 Ga ([Supplementary Fig. S3a, d–f](#)), 0.8 Ga ([Supplementary Fig. S3b](#)), or 0.6 Ga ([Supplementary Fig. S3c](#)). Assuming the transition between 2 lifestyles (states), host-associated and free living, next we used *phytools* v0.2-40 ([Revell 2024](#)) to simulate lifestyle evolution starting from a host-associated LCA of the bacterial clade under

3 different rate matrices, $\mathbf{Q} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ ([Supplementary Fig. S3a–c, 3f](#)), $\mathbf{Q} = \begin{pmatrix} -5 & 5 \\ 5 & -5 \end{pmatrix}$ ([Supplementary Fig. S3e](#)),

and $\mathbf{Q} = \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix}$ ([Supplementary Fig. S3f](#)). We per-

formed 20 ASR analyses to reconstruct the ancestral lifestyle of each internal node in the bacterial clade, and the result was averaged. The probability that the LCA of the bacterial clade takes a host-associated or a free-living lifestyle was later used in the pRTC framework whose divergence time was forced to be younger than that of the host with a certain probability (for details see the previous section). Furthermore, we employed *AliSim* ([Ly-Trong et al. 2022](#)) to simulate alignments of 200 amino acids under the substitution model LG + G ([Supplementary Fig. S3a–e](#)) or LG + G + C20 ([Supplementary Fig. S3f](#)) with an independent relaxed-clock model (see [Supplementary Note S4.1](#)). The above was repeated 30 times to generate 30 data sets under each of the 6 situations that is, the 6 panels in [Supplementary Fig. S3](#). For each data set, time estimation was performed both with (boxplots in green in [Supplementary Fig. S4](#)) and without pRTC (blue) to compare their results based on the difference in the time estimates from the true values (those in the simulated timetree). Note that, for simplicity, the lifestyle and sequences were simulated independently. However, in real cases, they may not be independent, which would be worth exploring in future studies.

Bayesian Sequential Approach to Dating the Bacterial Tree of Life

We applied a Bayesian sequential approach ([dos Reis et al. 2012](#); [Álvarez-Carretero et al. 2022](#)) to establish a

timeline of the bacterial tree, which allows using the genome-scale data and abundant fossil-based calibrations of eukaryotes to better constrain symbionts' ages through the pRTC framework (see [Supplementary Note S1](#) for details). The Bayesian sequential approach for clock-dating analysis consists of: 1) running a dating analysis of the first subset of data and collect the posterior samples of times and rates, 2) approximating the estimated posterior time densities in the first-step analysis by a parametric probability distribution and use it as the prior in a subsequent analysis with a second data subset with different genes from the first one (see [Supplementary Materials](#) for details). We first used the PAML dating program MCMCtree (PAML v4.10.0, [Yang 2007](#)) (see [Supplementary Note S1.2](#) for details) to date a eukaryote tree with 29 selected species based on our previous study ([Wang and Luo 2021](#)) and 320 orthologs ([Strasser et al. 2021](#)) with 11 fossil-based calibrations ([Supplementary Note S3.2](#)). Then, we fitted ST, SN, and Gamma (G) distributions to the posterior time densities inferred with this first data subset so that we could use them as priors to calibrate the matching eukaryotic nodes in the tree topology fixed during the second part of the analysis (i.e., bacterial tree with a eukaryotic subtree). Note that the 2 data subsets were non-overlapping (i.e., different genes between the first and the second data subsets).

Together with the fitted skew-t (ST), skew-normal (SN), and G distributions, we used the following bacterial calibrations to constrain the node ages in the second part of the Bayesian sequential dating approach ([Supplementary Note S3.2](#)). The calibrations from within bacteria were the 1.6 Ga-old akinete fossils for the filamentous cyanobacteria *Nostocales* ([Tomitani et al. 2006](#)), the 1.7 Ga-old microfossils for the coccoid cyanobacteria order *Pleurocapsales* ([Zhang and Golubic 1987](#)), the GOE (Great Oxidation Event) ~2.32 Ga for photosynthetic cyanobacteria ([Luo et al. 2016](#)), and the hydrocarbon biomarkers for green and purple sulphur bacteria both dated at 1.64 Ga ([Brocks et al. 2005](#)). We used the maximum age of the aforementioned calibrations, 2.32 Ga, to constrain the minimum age of the root. If not otherwise specified, in all molecular clock analyses we followed [Betts et al. 2018](#) to apply a uniform distribution as the prior: a soft tail was applied to the upper calibration bound, which means a small probability (by default 2.5%) where the age is beyond the bound, and a hard bound was set to further constrain minimum age, meaning that the age cannot be lower than the bound. This makes sense as minimum bounds are based on solid fossil or biomarker records while the maxima are often based on indirect evidence ([Benton and Donoghue 2007](#); [Marshall 2019](#)).

A Bootstrap-Based Approach to Improving MCMCtree Branch Length Estimate by Incorporating the Mixture Substitution Model

We used the approximate likelihood calculation implemented in MCMCtree ([dos Reis and Yang 2011](#))

to speed up timetree inference. Specifically, CODEML, will be called to first compute the branch lengths, the Hessian, and the gradient, which are then used by MCMCtree to approximate the likelihood calculation. Different from previous studies which often use the LG + G substitution model, we applied the bootstrap-based method developed in the present study to integrate the site-heterogeneous profile mixture model LG + G + C60 (Quang et al. 2008) implemented in IQ-Tree when calculating the branch lengths, the Hessian, and the gradient with CODEML. We based our method on the assumption that the maximum likelihood (ML) estimates of branch length follow a multivariate normal (MVN) distribution, according to the ML estimation theory stating that ML estimators are asymptotically multivariate normally distributed (Pawitan 2003). The idea was originally proposed by Thorne et al. 1998 and later implemented in PAML-MCMCtree as the approximate likelihood method (dos Reis and Yang 2011). This approach first obtains the maximum-likelihood estimates (MLEs) of the branch lengths $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)^T$, plus the gradient vector $\mathbf{g}$ and the Hessian matrix $\mathbf{H}$, which are the first and second derivatives of the log-likelihood evaluated at $\hat{\theta}$, respectively. Next, according to Eq. (1) in dos Reis and Yang (2011), the Taylor expansion at the second-order derivative was applied to approximate the log-likelihood

$$\ell(\theta) = \ell(\hat{\theta}) + \mathbf{g}^T \Delta \theta + \frac{1}{2} \Delta \theta^T \mathbf{H} \Delta \theta \quad (6)$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_n)^T$ denotes the parameters to estimate (branch lengths) and $\ell(\theta)$ is the log-likelihood as a function of branch lengths.

While the gradient is zero at the ML estimate ($\mathbf{g} = \mathbf{0}$), directly calculating $\mathbf{H}$ under IQ-Tree's profile mixture and other complex substitution model is difficult and has not yet been implemented in the program. However, note that when the sample size is large enough, the Hessian matrix evaluated at $\hat{\theta}$ is approximately the negative inverse matrix of the covariance matrix of the MLEs $\mathbf{V}(\hat{\theta})$ (Pawitan 2003):

$$\mathbf{H} \approx -\mathbf{V}(\hat{\theta})^{-1}. \quad (7)$$

Hence, we instead applied a bootstrap method to estimate $\mathbf{V}(\hat{\theta})$ by the bootstrap covariance matrix $\hat{\mathbf{V}}(\hat{\theta}^*)$ (Efron 1982) by taking B bootstrap alignments and estimating the MLEs of branch lengths $\hat{\theta}^*$:

$$\mathbf{V}(\hat{\theta}) \approx \hat{\mathbf{V}}(\hat{\theta}^*) = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}^*) (\hat{\theta}^{(b)} - \hat{\theta}^*)^T \quad (8)$$

where $\hat{\theta}^{(b)}$ indicate the MLEs of the branch lengths in the b th bootstrap sample. Specifically, 1000 bootstraps ($B = 1000$) were performed using the site-heterogeneous profile mixture model LG + G + Cxx (Cxx represents C10–C60) with a fixed tree topology by IQ-Tree, therefore generating 1000 bootstrap MLEs for each branch length and allowing calculating the bootstrap covariance matrix $\hat{\mathbf{V}}(\hat{\theta}^*)$. Accordingly, the

Hessian matrix can be approximated by solving the inverse matrix of $-\hat{\mathbf{V}}(\hat{\theta}^*)$. We obtained the MLEs of branch lengths, the gradient as a zero matrix, and the Hessian matrix approximated by the above procedure, and generated the "in.BV" file required by MCMCtree to approximate the likelihood calculation (dos Reis and Yang 2011). A series of simulations were performed to show the feasibility of this approach to integrating profile mixture substitution model into timetree inference analyses with MCMCtree (Supplementary Note S4).

Selection of the Best-Fitting Clock Model

We employed the mcmc3r package (dos Reis 2017; dos Reis et al. 2018) to determine the best-fitting clock models. This method can be used to find the best-fitting clock model by calculating the marginal likelihood using either the stepping-stone method (Xie et al. 2011) or the thermodynamic integration (Lartillot and Philippe 2006; Lepage et al. 2007). Specifically, we tested all the available clock models implemented in MCMCtree: the auto-correlated rate (AR; also known as the geometric Brownian motion), the independent-rate (IR) log-normal, and the strict (STR) clock models. Note that the exact likelihood calculation implemented in MCMCtree is required to collect the samples for each power posterior (McGowen et al. 2020) that the mcmc3r R package will later use to estimate the marginal likelihood with which Bayes factors and posterior probabilities can be computed. Note that the exact likelihood method implemented in mcmc3r works on only nucleotide sequences. We therefore recoded the amino acid sequences into 4-character states based on the Dayhoff-4 scheme (McGowen et al. 2020; Wang and Luo 2021) and proceeded to run MCMCtree and, subsequently, mcmc3r as aforementioned.

In the first-step sequential analysis, that is, dating the eukaryote tree based on the 320 orthologs identified in (Strasser et al. 2021), the IR and AR models were favored by similar proportions of genes (53% vs. 47%; *eukaryotic_tree* in Supplementary Fig. S5). Because posterior time estimates obtained by IR and AR were similar, we arbitrarily chose the IR model. For the second-step analysis, that is, dating the bacterial tree (148 species), we followed prior studies (McGowen et al. 2020; Wang and Luo 2021) to overcome the huge computational burden: 1) we compiled 2 reduced data sets by randomly selecting 20 and 40 species and 2) we fixed the root prior at 4.0 Ga and applied only a single internal calibration that constrained the age of the LCA of eukaryotes to 1.7–1.5 Ga (here, we did not want to estimate divergence times but simply select a better-fitting model). The AR model was shown to have the higher marginal likelihood for 69% and 73% of the genes for the 20- and 40-species sets (Supplementary Fig. S5), respectively, and was therefore selected as the preferred model, as done in a previous study (McGowen et al. 2020).

Selection of Genes With High Verticality or Low Among-Branch Rate Heterogeneity

To test if genes used in analysis impacted the posterior dates in molecular dating of the bacterial tree, the 19 mitochondrial genes were evaluated with the so-called ΔLL (difference in log-likelihood) as a metric of marker gene verticality (i.e., if genes were vertically inherited in evolution) (Moody et al. 2022), and relative rate difference which measures the extent of the violation of molecular clock for a gene (Wang and Luo 2021). Specifically, ΔLL calculates the difference in likelihood in log scale between a constrained ML tree based on the topology of the species tree and the ML gene tree. Genes with a higher ΔLL indicate large topology difference between the gene tree and the species tree, thus lower verticality (Moody et al. 2022). To this end, ΔLL well serves as a proxy for the extent to which the gene tree differs from the species tree. The relative difference in rate between mitochondria and others was calculated as $\frac{|r_{\text{mito}} - r_{\text{bac}}|}{\max(r_{\text{mito}}, r_{\text{bac}})}$ (Wang and Luo 2021), where the substitution rates of mitochondria or bacteria lineages were estimated by CODEML from PAML (Yang 2007). The higher the relative rate difference is, the more rate heterogeneity between lineages would be expected. Selecting only those with the smallest relative rate difference in molecular dating could ensure less violation of the molecular clock caused by the fast-evolving mitochondrial lineages, as used in Wang and Luo 2021. We selected the 5, 10, and 15 top-ranking genes using each of the above 2 measures and repeated the molecular clock analyses. Additionally, we performed the analysis with the full set of 32 mitochondria-related genes.

RESULTS

A New Framework for Dating Deep-Time Bacterial Evolution Informed by Ancient Bacterial Symbiosis

We propose a new framework to address the paucity of bacterial fossils in dating the bacterial tree by leveraging the temporal information of eukaryotes recorded in their abundant fossils. Briefly, by leveraging mitochondrial endosymbiosis which states that mitochondria of eukaryotes are phylogenetically affiliated with α -Proteobacteria (Martijn et al. 2018; Muñoz-Gómez et al. 2022), we incorporated eukaryotes into a bacterial phylogeny. Through this, it is tempting to co-estimate the ages of both the eukaryotic host and bacterial symbiont (Wang and Luo 2021) while applying the recently developed RTC-based molecular dating (Davín et al. 2018; Magnabosco et al. 2018; Szöllösi et al. 2022). In the context of host-symbiont association, this means constraining the origin time of the symbiont's crown group to be no earlier than that of the host's total group, thereby providing maximum time constraints to symbionts whose temporal information may further be propagated across bacteria.

It is well possible that the ancestor of modern symbionts adapted to a different host or even a free-living lifestyle. We formulate a 3-step approach accounting for the possibility of host shift of symbionts through ASR and fit it into RTC-based dating in a probabilistic framework, which we call probability-based RTC (pRTC; see Methods for details). First, ASR is performed to estimate the probabilities that the ancestral nodes of modern symbionts are associated with different hosts or are free-living. Second, Bayesian clock dating is performed using mitochondria-associated genes with fossil calibrations but without RTC to infer posterior time estimates where the ages of symbionts and eukaryotic hosts (placed as the mitochondrial clade sister to α -proteobacteria; see also above) are co-estimated. Third, the posterior timetrees collected in Step 2 are examined to remove those where the host's age is younger than the symbiont's, weighted by the probability of different states (hosts) taken by the ancestor of modern symbionts estimated in Step 1, using a rejection sampling approach. The idea is further illustrated in Fig. 1a. Assume that ASR infers the ancestral lifestyles of "symbiont1" as $P(S_1 = \text{mammal}) = 0.9$ and $P(S_1 = \text{freeliving}) = 0.1$, and for "symbiont2" as $P(S_2 = \text{mammal}) = 0.5$ and $P(S_2 = \text{freeliving}) = 0.5$. For the 2 timetrees in the posterior samples shown in Fig. 1a, the probabilities of acceptance are $(0.9 \times 1 + 0.1 \times 1) \times (0.5 \times 1 + 0.5 \times 1) = 1$, and $(0.9 \times 1 + 0.1 \times 1) \times (0.5 \times 0 + 0.5 \times 1) = 0.5$, respectively (see Eqs. 3–5 in Methods). If the above procedure is performed for all posterior timetrees collected from Step 2, the posterior time estimates of all bacterial lineages, particularly the symbiont, might shift toward the present due to the use of RTC informed by host-symbiont association (Fig. 1a). As ancient bacterial symbionts are distributed in different deep-branching lineages (Toft and Andersson 2010; McCutcheon and Moran 2012), this method has a great potential to help date the bacterial tree of life by including additional relatively ancient time constraints.

We further illustrate this idea by carrying out various simulations (see Methods; Supplementary Figs. S3 and S4). We simulated a species tree resembling the mito-endosymbiosis-based strategy mentioned above where the mitochondrial lineages (eukaryotes) are sister to α -proteobacteria (see Methods), and fixed the ages of the root and the LCA of the host at 2.0 Ga and 1.0 Ga, respectively. We simulated the sequence alignments, conducted ASR inference based on simulated lifestyle evolution, and used MCMCtree for timetree inference before and after pRTC. In all analyses, we showed that including pRTC (light green boxplots in Supplementary Fig. S4) significantly improved time estimates (paired *t*-tests; *P*-values < 0.001 in all comparisons; Supplementary Fig. S4). When the host clade was calibrated (boxplots on the right-hand side in Supplementary Fig. S4), the pattern became stronger, suggesting that estimating a more accurate divergence time for the host can improve divergence time estimation for the rest of the bacterial nodes (i.e., higher accuracy) within the pRTC framework.

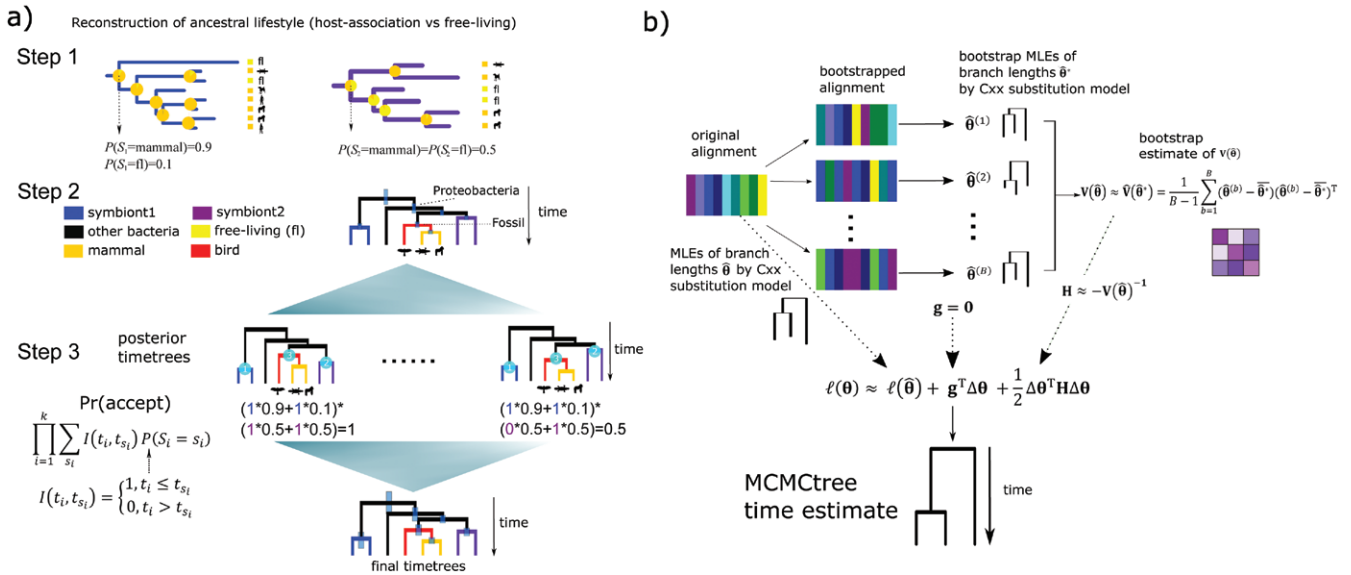


FIGURE 1. Graphical illustration of the 2 new methods in molecular dating developed in the present study. (A) Procedure of the pRTC dating informed by host-symbiont association. Suppose that there are 2 symbiotic lineages, Symbiont1 from non-Proteobacteria (blue) and Symbiont2 from α -proteobacteria (purple), and most of their members are mammal symbionts (orange) but some are free-living (yellow). First, ASR is performed on the phylogram of 16S rRNA tree to obtain the probability of each ancestral state taken by the LCA of the 2 “symbiont” clades respectively, that is, $P(S_1)$ and $P(S_2)$. Second, a phylogeny consisting of symbionts and mammals is constructed by mitochondrial genes, and MCMCtree is run with fossil calibrations but without RTCs. Third, rejection sampling is performed to remove posterior timetrees obtained in the previous step based on the requirement that symbiont’s origin must postdate the host. Specifically, we require that the age of the crown group of the symbionts (node1 and node2) to be no older than that of the total-group mammals (node3), weighted by $P(S_i = \text{mammal})$. The use of mammals’ total group is because there could be stem-group mammals that are unsampled or go extinct but that have Symbiont1 or Symbiont2 as symbiont. For the free-living state, we define $I(t_i, t_{fl}) = 1$ as the RTC is automatically satisfied for free-living bacteria (see Methods). In the final timetree, the divergence time of symbiont lineages becomes younger than that inferred before host association is considered, suggesting that incorporating RTC informed by host association improves the molecular clock inferences. The icons of animals are obtained from PhyloPic (<https://www.phylopic.org/>) under CC0 1.0 Universal. (B) Overflow of the bootstrap-based approach to incorporating branch length estimates inferred under site-heterogeneous profile mixture substitution models such as LG + G + Cxx where xx indicates 10–60 categories of equilibrium frequency. $\theta = (\theta_1, \theta_2, \dots, \theta_n)^T$ denotes the parameters to estimate (branch length, i.e., expected number of amino acid substitutions per site) and $\ell(\theta)$ is the log-likelihood as a function of branch lengths. Briefly, the MLE of branch length $\hat{\theta}$ is obtained by LG + G + Cxx, and the bootstrap covariance matrix of branch lengths $\mathbf{V}(\hat{\theta})$ is calculated to approximate the Hessian matrix (second-order derivative of the log-likelihood of branch length), which are further used for time estimation by MCMCtree.

When the “true” age of the symbiont was gradually reduced from 1.0 Ga to 0.6 Ga, making it progressively younger than the host, the improvement in time estimates provided by pRTC decreased (panels A, B, and C in [Supplementary Figs. S3 and S4](#)). Additionally, when the lifestyle was simulated to change more rapidly during evolution so that the phylogenetic signal was weakened, the improvement in time estimates became less pronounced, probably because the ASR inference is less accurate in such cases (panels A, D, and E in [Supplementary Figs. S3 and S4](#)).

Integrating Eukaryote Temporal Information by a Bayesian Sequential Approach

We reconstructed the phylogeny of bacteria using the 265 genomes and 60 gene markers carefully selected in a comprehensive phylogenomics study (Coleman et al. 2021) using the best-fitting site-heterogeneous profile mixture model LG + G + C60 ([Supplementary Table S1a](#) and [Supplementary Note S1](#)). We further applied the time-nonreversible model (Naser-Khdour et al.

2022) which infers the root of the bacterial tree as part of ML phylogenetic reconstruction where different root positions may have different likelihoods. This method does not depend on any outgroup, thereby avoiding long-branch attraction artifacts caused by using highly divergent sequences to root deep phylogenies. Our analyses rooted the bacterial phylogeny in the neighborhood of a clade consisting of Fusobacteriota and DST (Deinococcota, Synergistota, and Thermotogota). The analyses also supported the division of most bacterial lineages into Gracilicutes and Terrabacteria (Taib et al. 2020), and placed the candidate phyla radiation (CPR) (Patescibacteria) clade ([Supplementary Fig. S6](#)), initially suggested as the earliest-split bacterial group (Hug et al. 2016; Zhu et al. 2019), as a late-split clade sister to Chloroflexi ([Supplementary Fig. S6](#), [Supplementary Table S2](#), and [Supplementary Note S2](#)). This is generally consistent with a recent study that roots the bacterial tree using an independent method based on gene-species tree reconciliation (Coleman et al. 2021).

To apply pRTC based on host-symbiont co-evolution, we integrated eukaryotes and bacteria in a

unified framework to co-estimate their times using the mito-endosymbiosis-based dating strategy (Wang and Luo 2021). As noted, the ages of symbionts are more precisely and accurately estimated if the host date is better constrained, thus with less uncertainty. We therefore used the genome-scale data of eukaryotes by employing a Bayesian sequential approach in which the posterior times estimated with a data set with few taxa and many genes are estimated to serve as the time priors of a second analysis with many organisms but fewer genes (dos Reis et al. 2012; Álvarez-Carretero et al. 2022). Note that the Bayesian sequential clock-dating approach is conceptually distinct from secondary calibrations. In Bayesian sequential clock dating, the posterior time distributions obtained with the first data set are subsequently used as prior distributions with the second data set (fewer and different genes from the first data set but more taxa, including those taxa used in the first data set), which helps accommodate the uncertainty in divergence times estimation from 1 analysis to another. Here the 2 data sets must be independent (i.e., different genes in data set 1 and data set 2). In contrast, dating with secondary calibrations typically relies on point estimates of divergence times (though approximated time distributions can also be used) derived from other studies, and it often employs genes overlapping with those used to obtain the original calibrations from those studies (dos Reis et al. 2018). If the same genes are present in both data sets, the likelihood will be squared, and therefore its calculation during the MCMC will be wrong; hence the importance of independent data sets (Álvarez-Carretero et al. 2022) (Supplementary Note S1.2). We estimated the divergence times of 29 representative eukaryotes with 11 fossil calibrations (Supplementary Figs. S7 and S8) and 320 orthologs conserved across eukaryotes (Strasser et al. 2021). The posterior time estimates of eukaryote nodes were generally in agreement with most studies (Parfrey et al. 2011; Betts et al. 2018; Gibson et al. 2018; Wang and Luo 2021) (see Supplementary Note S5 for additional discussion). Next, we fitted ST, SN, or G distributions (Supplementary Fig. S9 and Supplementary Note S1.2.2) to the posterior time distributions inferred for each node in the fixed eukaryotic tree topology (first step of the Bayesian sequential clock-dating approach; “backbone” tree with 29 eukaryotic representative taxa and 320 orthologs) so that they could be used as priors to subsequently calibrate the eukaryotic nodes in the second data set with more taxa and less genes. In this second data set, the eukaryotic subtree was then placed as a sister clade to α -proteobacteria in the larger bacterial tree that was used for timetree inference during the second step of this Bayesian sequential clock-dating approach.

A Time-Calibrated Bacterial Phylogeny Based on Ancient Symbiosis

To apply the pRTC dating approach based on ancient symbiosis, we first performed comprehensive

MCMC-based ASR to infer the ancestral lifestyles of 7 well-known ancient clades where symbionts are widely distributed, 4 at a phylum level (Spirochaetota, Tenericutes, Chlamydiae, and Elusimicrobiota) and 3 at an order level (Rickettsiales, Holosporales, and Legionellales). The ASR was performed using the 16S rRNA gene, which best captures bacterial lifestyle diversity by including environmental and uncultured samples, based on the lifestyle information of modern members within each of the ancient bacterial clades mainly consisting of symbionts (Supplementary Figs. S10 and S11). We additionally included 2 genus-level symbionts (*Buchnera* and *Blattabacterium*) which are classical host-symbiont co-evolution examples (Perreau and Moran 2022). A total of 19 ancestral nodes consisting of 28 symbiont lineages from the above 9 symbiont clades were selected to represent the host-symbiont association-based RTCs in subsequent analyses (Fig. 2, Supplementary Fig. S10, and Supplementary Note S3.1).

Furthermore, we identified 19 genes well conserved across bacteria (Supplementary Figs. S7 and S8) from 108 genes evolutionarily associated with mitochondria (Muñoz-Gómez et al. 2022) (see also Supplementary Data S1). We performed molecular clock-dating analysis using a combination of these 19 genes and a 148-genome focal data set consisting of 29 eukaryotes, 28 symbionts, and 91 non-symbionts. The phylogeny was based on the tree reconstructed above using the nonreversible model and with eukaryotes placed as the sister to α -proteobacteria. The fitted ST, SN, and G distributions onto the posterior time densities inferred during the first step of the Bayesian sequential clock-dating approach, together with the 5 lineage-specific fossil- or biomarker-based calibrations within bacteria, were used to constrain various node ages in the bacterial tree, embedded with the eukaryotic subtree, during the second step of this timetree inference analysis. We used the approximate likelihood calculation implemented in MCMCTree (dos Reis and Yang 2011) to speed up the timetree inference given our large phylogenomic data set. However, the amino acid substitution models (e.g., LG + G) used to calculate branch lengths in PAML-MCMCTree (and many other molecular dating software such as BEAST and MrBayes) do not consider across-site compositional heterogeneity and often show poor model fit in reconstructing deep phylogenies (Gouy et al. 2015; Moody et al. 2022; Baños et al. 2024).

We therefore developed a bootstrap-based approach to incorporating branch length estimates inferred under the site-heterogeneous profile mixture substitution model LG + G + C60, which better accommodates issues in deep-time phylogenetics like substitution saturation and can be easily used by MCMCTree for timetree inference. Briefly, ML trees were built for 1000 bootstrap alignments, and the negative inverse of the bootstrap covariance matrix of their branch lengths, which can be calculated with either IQ-TREE or any other software implementing mixture models, was calculated

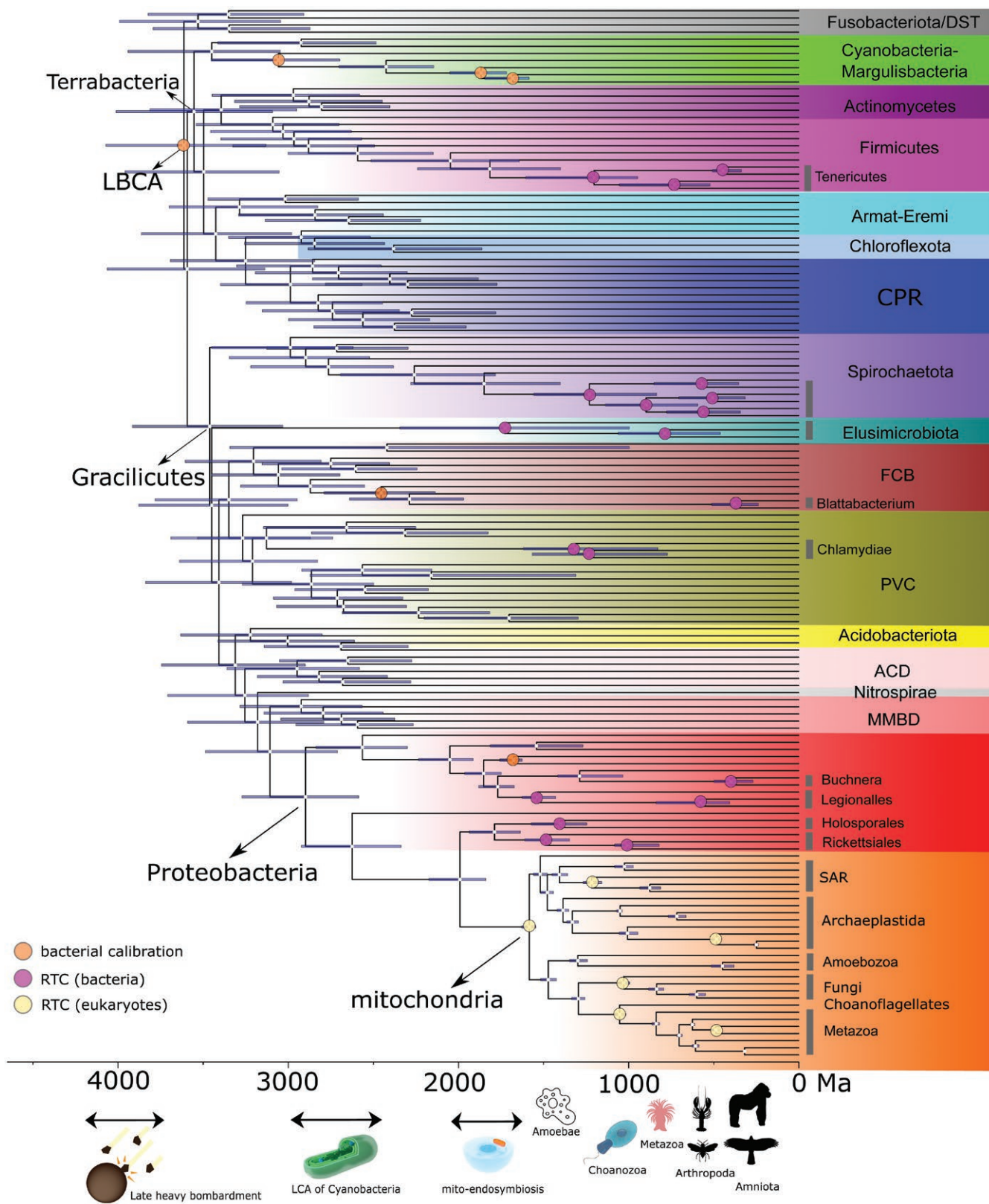


FIGURE 2. An evolutionary timeline of the bacterial tree of life established by the 3-step pRTC procedure. The root position is determined using IQ-Tree's nonreversible substitution model in ML phylogenomic tree reconstruction using the 19 mitochondrial genes. Node bars indicate the 95% HPD interval of age estimates. Nodes with circles indicate the bacterial calibration points (including the root). Eukaryotic calibrations obtained from sequential Bayesian dating are given in [Supplementary Fig. S4](#). RTCs of the 28 bacterial symbionts and 5 eukaryotic hosts (LCA

to approximate the Hessian matrix (which describes the curvature of the log-likelihood function around the MLEs of branch lengths); note that by default the branch lengths and Hessian matrix are calculated with CODEML (or BASEML for nucleotides) under only simpler substitution models like LG + G. Simulations showed considerable improvements in the estimates of both ages and absolute rates by integrating site-heterogeneous models with this approach, particularly in the absence of internal calibrations or if the divergence times were older than 3.0 Ga (Supplementary Figs. S12 and S13; see Supplementary Note S4 for details).

Next, we followed Eq. (5) (see Methods) to apply a rejection sampling approach to filter out the inferred timetrees without applying RTC from Step 2, based on the probabilities of the ancestor of modern symbionts being associated with different eukaryotic hosts or free-living inferred from the first step in the sequential dating (Fig. 1a and Supplementary Fig. S14c). Collectively, using a carefully selected focal dating scheme, we estimated that the LBCA occurred around 3648 Ma (95% highest posterior density [HPD]) 3162–4111 Ma) (Fig. 2), followed by the split between Terrabacteria and Gracilicutes estimated to be 3629 Ma (95% HPD 3167–4103 Ma). Most major clades, many of which are beyond a phylum level, diverged during 3000–3500 Ma (Fig. 2).

We gradually included eukaryotic temporal constraints, the LG + G + C60 substitution model, and host-symbiont association-based RTCs to check the improvement of posterior time estimates of the bacterial tree of life (lower triangle in Fig. 3a). The inclusion of eukaryote's timing (*Strategy2*: bacterial calibrations plus eukaryote time estimates by sequential molecular dating, substitution model LG + G, no RTCs) led to a $9.6 \pm 4.5\%$ (mean $\pm$ standard deviation) decrease in the posterior time estimates compared with *Strategy1* where the 5 bacterial calibrations were the only constraints (*Strategy1*: 5 bacterial calibrations with no eukaryote timing information, substitution model LG + G, no RTCs). When the branch lengths were estimated under the LG + G + C60 model (*Strategy3*: bacterial calibrations, posterior eukaryote time estimates with the sequential clock-dating approach, substitution model LG + G + C60, no RTCs), those nodes between 3000 and 2000 Ma were estimated to be older by $5.9 \pm 4.4\%$ (*Strategy2* vs. *Strategy3*). Furthermore, when pRTC was enabled under *Strategy3*, the estimated posterior times for almost all nodes were systematically younger ($8.0 \pm 5.5\%$, *Strategy3* vs. *Focal* in

Fig. 3a; *Focal*: bacterial calibrations, eukaryote time estimates by sequential molecular dating, substitution model LG + G + C60, with RTCs), suggesting that the constraint on the time order imposed by host-symbiont co-evolution was propagated through the bacterial tree of life. Collectively, the combination of the use of time constraints provided by eukaryote temporal information, site-heterogeneous profile mixture models, and host-symbiont association-based RTCs estimated the divergence times of analyzed bacteria to be $14.0 \pm 7.5\%$ younger than those obtained without any of these settings (*Strategy1* vs. *Focal* in Fig. 3a [upper triangle]; please see Supplementary Note S1.2.1 for details on how to estimate the absolute rates; this represents a net reduction and some effects canceled out as illustrated above). As to the posterior rate for each branch, the use of C60 led to an almost doubling of rate estimate (upper triangle in Fig. 3a; *Strategy2* vs. *Strategy3*), indicating that the absolute rates are largely underestimated by the compositionally homogeneous LG + G model. Moreover, when we successively increased the maximum time prior of the root from 4500 Ma to 6000 Ma, the inferred posterior times of bacterial lineages were $\sim 20\%$ older under *Strategy1*, but remained almost unchanged with the focal, *Strategy2*, and *Strategy3* schemes (Fig. 3b and Supplementary Fig. S15a), indicating that the time constraints imposed by eukaryotes effectively constrained the timing of bacterial evolution, such that their time estimates were robust to the root calibration where much debate has arisen (Nisbet and Sleep 2001; Moody et al. 2022). The time estimates obtained by the focal scheme were consistently younger than those estimated by *Strategy3* (Fig. 3b), suggesting that the pRTC effectively constrained the divergence times.

Accommodating the Uncertainty in Divergence Time Estimation

We evaluated the impact that multiple sources of uncertainty can have on species' divergence times estimation. Firstly, fossil-based calibrations are known to play a major role in molecular dating (Ho and Duchêne 2014). We used a total of 17 calibrations: 11 to constrain node ages in the eukaryotic clades, 5 within bacterial clades, and a last calibration to constrain the root age. Replacing the 1.047 Ga-old fossil of *Bangiomorpha*, the most widely recognized oldest crown-group eukaryote (Betts et al. 2018; Gibson et al. 2018), with the recently discovered but disputed 1.6 Ga-old *Rafatazmia* (Bengtson et al. 2017) (see also Supplementary Note

of eukaryotes, total-group animals, total-group ciliates, total-group insects, and crown-group fungi) are denoted by circles. Because the 28 symbionts are selected from "symbiont" clades, a total of $28 - 9 = 19$ internal nodes serve as the RTCs for bacteria. The organism images are placed at the origin times of eukaryotes of interest co-estimated with bacterial lineages. Detailed posterior ages for clades of interest are summarized in Supplementary Table S4. DST: Deinococcota, Synergistota, and Thermotogota; Armat-Eremi: Armatimonadetes and Eremiobacteraeota. CPR: Candidate phyla radiation; FCB: Fibrobacterota, Chlorobiota, Bacteroidota; PVC: Planctomycetota, Verrucomicrobiota, Chlamydia, and related lineages; ACD: Aquificota, Campylobacterota, and Deferribacterota. MBDD (δ -Proteobacteria): Myxococcota, Bdellovibrionota, Desulfomonadota, and Desulfobacterota. The images of the choanozoan, and the cyanobacteria and mitochondria are credited to Urutseg and Kevin Song, respectively under CC BY-SA 3.0 (<https://creativecommons.org/licenses/by-sa/3.0/>) with slight modifications. All other organism images are distributed under CC 1.0.

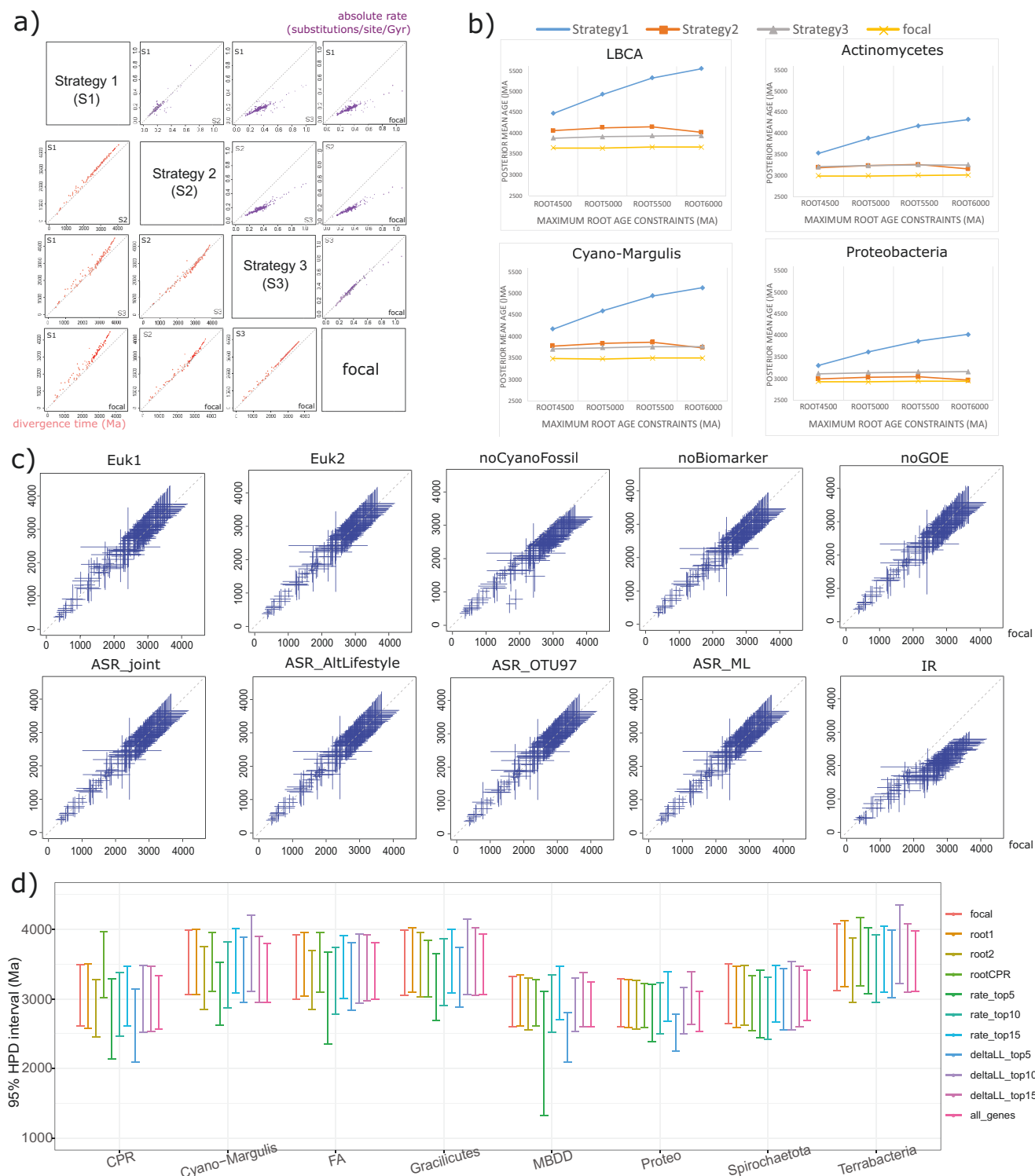


FIGURE 3. Assessing the uncertainty in Bayesian clock-dating analysis. (A) Comparison of the posterior mean ages (lower triangle; unit: Ma) and rates (upper triangle; unit: number of substitutions per site per Ga) of bacteria estimated with different strategies. *Strategy1* (traditional strategy used in prior studies): 5 bacterial calibrations with hard minimum bounds and a soft maximum bound < 4.5 Ga at LBCA, no eukaryote timing information, substitution model LG + G, no RTCs. *Strategy2*: bacterial calibrations, eukaryote times by Bayesian sequential clock-dating approach, substitution model LG + G, no RTCs. *Strategy3*: bacterial calibrations, eukaryote times by Bayesian sequential clock-dating approach, substitution model LG + G + C60, no RTCs. Focal: bacterial calibrations, eukaryote times by Bayesian sequential clock-dating approach, substitution model LG + G + C60, with RTCs. (B) Posterior mean ages of the selected clades calculated under root maximum prior ages from 4500 to 6000 Ma with the focal strategy and Strategies 1–3. (C) Divergence times of bacteria estimated by alternative schemes (alternative calibration hypotheses constraining differently the age of contentious nodes) (*y*-axis; [Supplementary Data S3](#)) versus the one in the focal

S3.2), pushed the LCA of eukaryotes from ~1600 Ma to ~2100 Ma (Supplementary Fig. S8), but had little impact on the time estimates of bacteria (*Euk1* in Fig. 3c and Supplementary Data S2). A similar pattern was observed when the recently discovered but disputed 0.89 Ga-old sponge fossil instead of the 0.55 Ga-old one (Supplementary Note S3.2) was alternatively used as the minimum age of the total-group animals (*Euk2* in Fig. 3c). Applying other eukaryotes' calibrations also showed very little changes to the posterior date estimates in the bacterial timetree (*Euk3-Euk5* in Supplementary Fig. S15 and Supplementary Data S3). Regarding the 5 bacterial calibrations, the 2 cyanobacteria fossils and the 2 biomarkers displayed considerable impacts on the posterior divergence time estimates as their removal led to a younger time estimate (7% on average) (*noCyanoFossil* and *noBiomarker* [y-axis] compared with the focal strategy [x-axis] in Fig. 3c and Supplementary Data S3). Replacing the 1.6 Ga-old Nostocales fossil calibration (Tomitani et al. 2006) by a more disputable estimate based on the 2.0 Ga-old fossil (Amard and Bertrand-Sarfati 1997) led to older posterior time estimates by more than ~10% (Supplementary Fig. S12b and Supplementary Data S3). Except the above, the use of all alternative fossil calibrations returned similar time estimates of both eukaryotes and bacteria (Supplementary Figs. S15b and S16).

Secondly, another source of time constraint is the use of ASR-based RTC. In the above analysis, the ancestral states (hosts) between nodes within the same tree were considered independent of each other such that their joint probability can be approximated by the multiplied marginal probability (Supplementary Fig. S2b). To check if this approximation affected the final time estimates, we directly estimated the joint probability of their ancestral states using an SCM method (see Methods). This generally did not change the posterior time estimates (*joint_prob* in Fig. 3c). Further, while we have conducted comprehensive work in classifying the traits into different hosts or free-living, there exist different interpretations of the phenotypic traits, particularly free-living, which might affect ASR. Analysis with alternative host/lifestyle classification (Supplementary Note S3.1.3) suggested little changes in posterior time estimates (*ASR_AltLifestyle* in Fig. 3c). Additionally,

we tried different cutoffs of 16S OTU and a different algorithm (ML) in ASR, all of which resulted in highly consistent time estimates (Fig. 3c). Including 13 more symbiont lineages returned similar time estimates (*more_RTCs* in Supplementary Fig. S15), suggesting that adding host association-based RTCs would not lead to further change in the posterior ages of bacteria.

Thirdly, dependence of the results on the phylogenetic tree topology was investigated. When timetree inference analysis was performed with 3 alternative root positions, 2 near the root (*root0*) used in the focal analysis (*root1* and *root2* in Fig. 3d and Supplementary Fig. S6), and 1 (*rootCPR*) at the CPR clade as suggested by earlier studies (Hug et al. 2016; Zhu et al. 2019), the posterior time estimates remained similar for most clades, except those that were alternatively placed as an early-split phylogenetic position, particularly the CPR. Alternative phylogenetic positions of mitochondria or between eukaryotic groups (Supplementary Fig. S14d,e) displayed minor impacts on the posterior time estimates of bacteria (Supplementary Fig. S15c).

Next, the use of different genes in dating may lead to different time estimates. Mitochondrial genes tend to evolve faster than their orthologs in bacterial genomes (Martijn et al. 2018; Wang and Luo 2021). In theory, the issue can be accommodated by relaxed molecular clock models, but this is not guaranteed (Drummond et al. 2006; Yang and Rannala 2006). Also, MCMCtree cannot determine which tree it should take. Instead, it takes a fixed tree topology during timetree inference, and so it does not integrate over alternative tree topologies that may be possible due to HGT, unresolved paralogy, or limited phylogenetic signal of individual genes. Nevertheless, if there are contested tree hypotheses that could be tested for a given phylogeny, it is possible to run a Bayesian model selection analysis with both MCMCtree and the R package *mcmc3r* (dos Reis et al. 2018). To minimize the bias caused by gene selection, we ranked the 19 mitochondrial genes by 1) the relative difference in their estimated evolutionary rates between eukaryotic and bacterial lineages and 2) ΔLL , an index to measure the species- and gene-tree incongruence (Moody et al. 2022) (see Methods). We selected the 5, 10, and 15 top-ranking (best) genes using each of the above 2 strategies and repeated clock-dating analyses

strategy (x-axis). The bars denote the 95% HPD. *Euk1*: the minimum time bound of crown-group red algae alternatively set as 1.6 Ga based on *Rafatatzmia*; *Euk2*: the minimum time bound of animal crown group alternatively set as 0.89 Ga (Supplementary Note S3.2); *noCyanoFossil*: the 2 cyanobacteria minimum time bounds (total-group Nostocales and Pleurocapsales) removed; *noBiomarker*: the 2 bacterial biomarker minimum time bounds (total-group Chromatiaceae and Chlorobi) removed; *noGOE*: the minimum time bound based on GOE removed. *Joint_prob*: rejection sampling based on calculating the joint probability (see Supplementary Fig. S6b) in Eq. (2). *Alt_lifestyle*: ASR inferred with alternative classification of lifestyles of modern symbionts (Supplementary Note S3.1). *OTU97*: ASR performed with OTU at a cutoff of 97% sequence identity. *ASR_ML*: maximum likelihood algorithm instead of MCMC used in ASR. *IR*: the independent rate model is used instead of the AR model. (D) Changes in the posterior time estimates shown as 95% HPD interval resulting from different root positions of the phylogenomics tree (*root1*, *root2*, and *rootCPR*), and the use of different genes. *rate_top* and *deltaLL_top*: the 5, 10, and 15 (best) top-ranking genes according to relative rate difference between mitochondrial and bacterial lineages or ΔLL which measures the degree of species- and gene-tree incongruence with *root0* as the root position (see Methods and Supplementary Data S3). *all_genes*: molecular dating performed on all 32 mitochondrial genes conserved across bacteria (also *root0* as the root position). The center is the posterior mean age. FA: Firmicutes and Actinomycetes. MBDD (δ -Proteobacteria): Myxococcota, Bdellovibrionota, Desulfomonadota, and Desulfobacterota.

(Fig. 3d). Additionally, we analyzed the full set of 32 mitochondrial genes (*all_genes* in Fig. 3d). Most analyses showed consistent results compared with using the 19-gene data set; the schemes using only the 5 top-ranking genes (schemes *top5*) displayed larger uncertainty in the date estimates of certain nodes (e.g., MBDD [Myxococcota, Bdellovibrionota, Desulfomonadota, and Desulfobacterota]), but this is likely due to their small number of genes.

Lastly, other factors potentially affecting the dating analysis were examined. The largest change in posterior time estimates was observed when the ARs model, which was consistently supported as the preferred model in the Bayesian model selection analysis with *mcmc3r* (Supplementary Fig. S5 and Supplementary Data S4), was replaced with the poorly supported IR model: the divergence times of most lineages shifted toward the present by ~20% (IR in Fig. 3c). The birth-death process with species sampling (Yang and Rannala 1997) is used to specify the prior on those nodes whose age is not constrained by any calibration (i.e., uncalibrated nodes). To generate the joint distribution of all node ages, both this process and the priors used as calibrations will be considered, which may result in a complicated joint prior. Additionally, the impact of changing the parameters of the birth-death process with species sampling was not significant (*bd* in Supplementary Fig. S15b). In addition, analysis with an independently sampled and expanded data set with 285 bacterial genomes (Coleman et al. 2021) showed similar posterior time estimates (Secondary in Supplementary Fig. S16), suggesting that the time estimates presented in the main analysis were not dependent on a specific taxon sampling. Note that the overall acceptance rate of the rejection sampling was less than 5%, meaning that more than 95% of the posterior samples were removed in Step 3 of the pRTC framework (see Methods). Because the running time of MCMCtree is roughly proportional to the number of partitions, to save time, all analyses were performed on a single partitioned alignment, but similar patterns were found when 2 partitions were used (Supplementary Fig. S17). In our timetree inference analysis, the calibration densities (i.e., node age constraints we specified, see Supplementary Note S1.2.2) differed from the marginal densities to be used by the dating program (also known as “effective priors”; solid vs. dashed lines in Supplementary Fig. S18 [focal strategy]) due to the overlap between various node age constraints we specified. This truncation issue is common in deep phylogeny dating given that the root age may be constrained with an arbitrary maximum age due to the uncertainty surrounding very ancient events. To alleviate the truncation effect, we assessed the impact of 3 additional sets of calibrations with less overlapping constraints, which resulted in more similar calibrations and marginal densities. The resulting posterior time estimates were very close to those obtained when using the “focal” strategy (Supplementary Fig. S18).

The convergence of all MCMC analyses was checked by comparing the time estimates of 2 independent chains (Supplementary Fig. S19) and by calculating the effective sample size (ESS) to ensure that the ESS of almost all parameters after rejection sampling was above 200 for any single chain in all MCMCtree analyses (Supplementary Data S5).

DISCUSSION

A major challenge in dating bacterial evolution is the paucity of credible calibrations, particularly the maximum time constraints (Wang and Luo 2021). Previously, this was usually done by assigning a somewhat arbitrary maximum age to the root (LBCA), which greatly varies between studies due to the uncertainty in palaeobiological evidence and phylogenetic relationships around the origin of life (Knoll 2015; Moody et al. 2022). We addressed this problem by leveraging the timing information of eukaryotes based on ancient symbiosis. This provides 2 additional layers of timing information besides bacterial calibrations. Firstly, we linked the evolution of bacteria and eukaryotes by mitochondrial endosymbiosis. Several eukaryotic calibrations incorporate upper bounds to establish a maximum age based on the most recent geological formation that ought to have fossils of the clade of interest. Consequently, these more reliable maximum ages can help inform bacterial evolution if eukaryotic lineages are incorporated into the tree topology. Secondly, based on RTCs informed by host-symbiont association, the timing of eukaryote evolution directly constrains the origin time of symbionts to be no earlier than their eukaryotic host, which is further propagated across the bacterial phylogeny.

The idea of RTC in molecular dating has recently caught much attention (Davín et al. 2018; Magnabosco et al. 2018; Fournier et al. 2021). However, to the best of our knowledge, none of the available RTC applications has considered the uncertainty in the inference of RTC. In other words, previously, the relative time order established by HGT (i.e., gene recipient lineage cannot be older than gene donor lineage) or any other forms of RTC was assumed to be known with certainty, which might be oversimplified due to the uncertainties in identifying these evolutionary events. Our new approach removes this assumption by weighting the different ancestral hosts by their probability established in ASR through a rejection sampling framework. This greatly extends the RTCs based on host-symbiont co-evolution to dating the bacterial tree, which involves many ancient symbioses and large uncertainty in assigning the ancestral host. In addition, prior studies that have applied the idea of host-symbiont association in molecular dating typically used time estimates of the host from other studies (i.e., secondary calibrations) to calibrate symbiont evolution, which may cause wrong time estimates by propagating errors associated with the original time estimates (Schenk 2016). Our strategy co-estimates the origin times of hosts and symbionts

based on mitochondrial endosymbiosis (Wang and Luo 2021), avoiding methodological and data differences in comparing posterior time estimates from different studies.

Great care should be taken when interpreting the estimated timeline of the bacterial tree of life and reconstructed evolutionary events taking place anciently. One notable issue is that reconstructing the ancestral lifestyle of bacteria is a very difficult task so there is still much uncertainty associated with the inference (Dharamshi et al. 2020, 2023). Accordingly, although we have spent many efforts minimizing this uncertainty (Fig. 3c, Supplementary Figs. S5 and S10, and Supplementary Note S3.1), the ASR results could still have some impacts on the estimated timetree. Nonetheless, it is remarkable to note the congruence between time estimates obtained using different time constraints, different root positions, independent sets of genomes, different parameters in molecular clock analysis, and, particularly, different settings in ASR which could lead to different ancestral lifestyles, and thus different RTCs. Our time estimate of LBCA is ~0.5 Gyr (billion years) younger than estimated in some studies (Battistuzzi and Hedges 2009; Marin et al. 2017) but is consistent with a recent study using modern phylogenetics software and carefully selected HGT-based time constraints (Fournier et al. 2021). As evident in Fig. 3a, this might be due to the lack of enough maximum time constraints and the use of poorly fit substitution models for inferring deep phylogenies in those early efforts. Moreover, we conducted molecular clock inferences based on a phylogeny of the bacterial tree of life constructed with state-of-the-art phylogenomics methods. Consequently, some groups once thought basal in the bacterial tree, for example, Aquificae (Battistuzzi and Hedges 2009) and CPR (Hug et al. 2016), are suggested as later-split lineages in recent studies (Coleman et al. 2021; Moody et al. 2022) and the present study based on improved phylogenetics methods, leading to different time estimates of corresponding lineages.

In addition, our model assumes the ancestral bacterium took only 1 host or lifestyle at 1 time. While it is straightforward to assume so, it is possible that some symbionts adapt to different lifestyles at different life stages or under different environmental conditions. Further, while bacterial symbionts isolated from animals are typically classified as animal-associated, some might actually be symbionts of the protists that live in the gut or other parts of the animals (Collingro et al. 2020). Additionally, some bacteria found in animal organs like the gut may be transient free-living bacteria, as these kinds of organs are connected to external environments. In sum, misclassification of bacterial lifestyles may indeed occur, which could affect ASR. Nonetheless, note that the internal nodes that served as the RTCs were selected from classical “symbiont clades” based on not only ASR but also a comprehensive literature review (Supplementary Note S3.1). Furthermore, the consistent results

obtained when alternative ways of lifestyle classification and parameters in ASR were used suggest that the above caveats likely did not have a large impact on the posterior date estimate in the RTC framework. This is likely because of the many RTCs included such that any change in a few of them is unlikely affected the overall pattern.

The estimated timescale of bacterial evolution allows testing hypotheses involving early life evolution. For example, it is generally believed that the late heavy bombardment likely occurred on early Earth ~4.1–3.8 Ga, during which large asteroids intensely pummeled Earth, melted most parts of its surface, and killed off any forms of life that might have begun to emerge (Abramov and Mojzsis 2009; Boehnke and Harrison 2016). Because the primordial form of life likely had existed before LBCA, our estimates for the age of LBCA at 4.0–3.5 Ga hint that life had survived the cataclysm, consistent with modeling studies showing that much of Earth’s crust and the microbes living therein were unlikely affected by the intense bombardment (Abramov and Mojzsis 2009). Under the most parsimonious hypothesis that oxygenic cyanobacteria arose in the cyanobacterial stem lineage (Sánchez-Baracaldo and Cardona 2020; Fournier et al. 2021), our analysis placed the appearance of the first oxygenic cyanobacteria between 3.0 Ga (total-group cyanobacteria) and 2.5 Ga (crown-group cyanobacteria) (Fig. 2). This reinforces the previous argument (Hofmann et al. 1999; Allwood et al. 2006) that one cannot link stromatolites found in ancient sediments dated ~3.5 Ga to oxygenic cyanobacteria despite their dominance in stromatolites nowadays. This is supported by recent evidence showing that stromatolites preserved in Archean rocks neither required oxygen-producing cyanobacteria for formation (Schopf 2011) nor were necessarily biogenic (McLoughlin et al. 2008). Additionally, our approach effectively constrained the total group and crown group of oxygenic cyanobacteria (Figs. 2 and 3d). These ages were controversial in previous studies using either cyanobacteria fossils or a combination of both cyanobacteria and plant fossils by a plastid-endosymbiosis strategy (Sánchez-Baracaldo et al. 2014, 2017; Fournier et al. 2021; Zhang et al. 2023).

The new bacterial timetree also makes it possible to infer the mechanisms that drive the origin of enzymes that involve oxygen. Accumulating evidence from isotope analyses suggests the presence of oxygen on Earth back to 3.1 Ga (Crowe et al. 2013; Planavsky et al. 2014; Wang et al. 2018), which might be produced by cyanobacteria as mentioned above, and/or abiotically from hydrogen peroxide (H_2O_2) (He et al. 2021), potentially leading to the formation of “oxygen oasis” thus extracellular access to O_2 . A recent study (Jabłońska and Tawfik 2021) tracked the origin of most oxygen-utilizing and -producing enzyme families to the Gracilicutes-Terrabacteria split, which according to our estimate is ~3.6 Ga, predating the above events for ~0.5 Ga (Fig. 2). This suggests that life had access to oxygen

earlier than the GOE (Jabłońska and Tawfik 2021), either extra- or intracellularly (e.g., oxygen production through nitric oxide dismutation in Thaumarchaea [Kraft et al. 2022] and the NC10 group [Ettwig et al. 2010]) but more likely the latter.

Our work shows that there is much more information beyond bacterial fossils for calibrating bacterial evolution. The estimated timeline of bacterial evolution may well serve as a foundation for studying the evolution of bacteria, their diversification rate, and their correlation with Earth's geochemistry. It is hopeful that some of these questions can be answered, and competing hypotheses can be reconciled by future molecular clock studies in combination with modern phylogenetics and comparative genomics methods.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: <https://dx.doi.org/10.5061/dryad.1c59zw42s>

ACKNOWLEDGEMENTS

We sincerely appreciate the detailed and constructive comments provided by the 2 reviewers. We are particularly grateful to Jianhao Lv for the critical comments on the mathematical derivation of the formulae, and the professional insights in statistics. We are also very grateful to Sandra Álvarez-Carretero and Edmund Moody for the efforts in testing the script *bs_inBV* and making it more user-friendly. We also thank Daniel Mills and Philip Donoghue for insights in the interpretation of eukaryotic fossils, Mario dos Reis, José Barba-Montoya and Suha Naser for help in phylogenetics analysis, and Andrew Meade, Stilianos Louca and Liam Revell for guidance in ancestral state reconstruction. We thank László Nagy, Torda Varga and Tomáš Flouris for kindly sharing with us software. We also thank Luo Lab members Tianhua Liao and Hao Zhang for discussion, and PhyloPic (<https://www.phylopic.org/>) for the free images of organisms.

FUNDING

The work is supported by the Hong Kong Research Grants Council General Research Fund (14116922), the Natural Science Foundation of China (42293294), and the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY

The scripts to do pRTC and bootstrap-based approach Hessian approximation for MCMCtree are available at <https://github.com/evolbeginner/rrtc> and https://github.com/evolbeginner/bs_inBV/ respectively. Other scripts and data are available at <https://doi.org/10.5061/dryad.1c59zw42s>.

REFERENCES

- Abramov O., Mojzsis S.J. 2009. Microbial habitability of the Hadean Earth during the late heavy bombardment. *Nature* 459:419–422.
- Allwood A.C., Walter M.R., Kamber B.S., Marshall C.P., Burch I.W. 2006. Stromatolite reef from the Early Archaean era of Australia. *Nature* 441:714–718.
- Álvarez-Carretero S., Tamuri A.U., Battini M., Nascimento F.F., Carlisle E., Asher R.J., Yang Z., Donoghue P.C.J., dos Reis M. 2022. A species-level timeline of mammal evolution integrating phylogenomic data. *Nature* 602:263–267.
- Amard B., Bertrand-Sarfati J. 1997. Microfossils in 2000 Ma old cherty stromatolites of the Franceville Group, Gabon. *Precambrian Res.* 81:197–221.
- Baños H., Susko E., Roger A.J. 2024. Is over-parameterization a problem for profile mixture models? *Syst. Biol.* 73:53–75.
- Battistuzzi F.U., Feijao A., Hedges S.B. 2004. A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol. Biol.* 4:44.
- Battistuzzi F.U., Hedges S.B. 2009. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* 26:335–343.
- Bengtson S., Sallstedt T., Belivanova V., Whitehouse M. 2017. Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLoS Biol.* 15:e2000735.
- Benton M.J., Donoghue P.C.J. 2007. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* 24:26–53.
- 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. *Nat. Ecol. Evol.* 2:1556–1562.
- Boehnke P., Harrison T.M. 2016. Illusory late heavy bombardments. *Proc. Natl. Acad. Sci. U.S.A.* 113:10802–10806.
- Brocks J.J., Love G.D., Summons R.E., Knoll A.H., Logan G.A., Bowden S.A. 2005. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* 437:866–870.
- Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- Coleman G.A., Davín A.A., Mahendrarajah T.A., Szánthó L.L., Spang A., Hugenholtz P., Szöllsi G.J., Williams T.A. 2021. A rooted phylogeny resolves early bacterial evolution. *Science* 372:eabe0511.
- Collingro A., Köstlbacher S., Horn M. 2020. Chlamydiae in the environment. *Trends Microbiol.* 28:877–888.
- Crowe S.A., Dössing L.N., Beukes N.J., Bau M., Kruger S.J., Frei R., Canfield D.E. 2013. Atmospheric oxygenation three billion years ago. *Nature* 501:535–538.
- Davín A.A., Tannier E., Williams T.A., Boussau B., Daubin V., Szöllsi G.J. 2018. Gene transfers can date the tree of life. *Nat. Ecol. Evol.* 2:904–909.
- Dharamshi J.E., Köstlbacher S., Schön M.E., Collingro A., Ettema T.J.G., Horn M. 2023. Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat. Microbiol.* 8:40–54.
- Dharamshi J.E., Tamarit D., Eme L., Stairs C.W., Martijn J., Homa F., Jørgensen S.L., Spang A., Ettema T.J.G. 2020. Marine sediments illuminate Chlamydiae diversity and evolution. *Curr. Biol.* 30:1032–1048.e7.
- dos Reis M. 2017. Estimating the marginal likelihood of a relaxed-clock model with MCMCTree. Available from: <https://dosreis-lab.github.io/2017/10/24/marginal-likelihood-mcmc3r.html>. Accessed January 2023.

- dos Reis M.D., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* 28:2161–2172.
- dos Reis M., Gunnell G.F., Barba-Montoya J., Wilkins A., Yang Z., Yoder A.D. 2018. Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: primates as a test case. *Syst. Biol.* 67:594–615.
- dos Reis M., Inoue J., Hasegawa M., Asher R.J., Donoghue P.C.J., Yang Z. 2012. Phylogenomic data sets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. Biol. Sci.* 279:3491–3500.
- Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88–710.
- Efron B. 1982. The bootstrap. In: *The Jackknife, the Bootstrap and Other Resampling Plans*, Vol. 38. Philadelphia, USA: SIAM. p. 27–36.
- Ettwig K.F., Butler M.K., Le Paslier D., Pelletier E., Mangenot S., Kuypers M.M.M., Schreiber F., Dutilh B.E., Zedelius J., De Beer D., Gloerich J., Wessels H.J.C.T., Van Alen T., Luesken F., Wu M.L., Van De Pas-Schoonen K.T., Op Den Camp H.J.M., Janssen-Megens E.M., Francois K.J., Stunnenberg H., Weissenbach J., Jetten M.S.M., Strous M. 2010. Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* 464:543–548.
- Falkowski P.G., Fenchel T., Delong E.F. 2008. The microbial engines that drive earth's biogeochemical cycles. *Science* 320:1034–1039.
- Fournier G.P., Moore K.R., Rangel L.T., Payette J.G., Momper L., Bosak T. 2021. The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. *Proc. Biol. Sci.* 288:20210675.
- Gibson T.M., Shih P.M., Cumming V.M., Fischer W.W., Crockford P.W., Hodgskiss M.S.W., Wörndle S., Creaser R.A., Rainbird R.H., Skulski T.M., Halverson G.P. 2018. Precise age of *Bangiomorpha pubescens* dates the origin of eukaryotic photosynthesis. *Geology* 46:135–138.
- Gouy R., Baurain D., Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140329.
- He H., Wu X., Xian H., Zhu J., Yang Y., Lv Y., Li Y., Konhauser K.O. 2021. An abiotic source of Archean hydrogen peroxide and oxygen that pre-dates oxygenic photosynthesis. *Nat. Commun.* 12:6611.
- Ho S.Y.W., Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* 23:5947–5965.
- Hofmann H.J., Grey K., Hickman A.H., Thorpe R.I. 1999. Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia. *Bull. Geol. Soc. Am.* 111:1256–1262.
- Huelsbeck J.P., Nielsen R., Bollback J.P. 2003. Stochastic mapping of morphological characters. *Syst. Biol.* 52:131–158.
- Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., Banfield J.F. 2016. A new view of the tree of life. *Nat. Microbiol.* 1:16048.
- Jabłońska J., Tawfik D.S. 2021. The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation. *Nat. Ecol. Evol.* 5:442–448.
- Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.
- Knoll A.H. 2015. Paleobiological perspectives on early microbial evolution. *Cold Spring Harb. Perspect. Biol.* 7:a018093–a018017.
- Kraft B., Jehmlich N., Larsen M., Bristow L.A., Könneke M., Thamdrup B., Canfield D.E. 2022. Oxygen and nitrogen production by an ammonia-oxidizing archaeon. *Science* 375:97–100.
- Lagesen K., Hallin P., Rødland E.A., Stærfeldt H.H., Rognes T., Ussery D.W. 2007. Rfam: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35:3100–3108.
- Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst. Biol.* 55:195–207.
- Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* 24:2669–2680.
- Li W., Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
- Liao T., Wang S., Stüeken E.E., Luo H. 2022. Phylogenomic evidence for the origin of obligate anaerobic Anammox bacteria around the great oxidation event. *Mol. Biol. Evol.* 39:msac170.
- Liao T., Wang S., Zhang H., Stüeken E.E., Luo H. 2024. Dating ammonia-oxidizing bacteria with abundant eukaryotic fossils. *Mol. Biol. Evol.* 41:msae096.
- Louca S., Shih P.M., Pennell M.W., Fischer W.W., Parfrey L.W., Doebeli M. 2018. Bacterial diversification through geological time. *Nat. Ecol. Evol.* 2:1458–1467.
- Luo G., Ono S., Beukes N.J., Wang D.T., Xie S., Summons R.E. 2016. Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. *Sci. Adv.* 2:e1600134.
- Ly-Trong N., Naser-Khdour S., Lanfear R., Minh B.Q. 2022. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Mol. Biol. Evol.* 39:msac092.
- Magnabosco C., Moore K.R., Wolfe J.M., Fournier G.P. 2018. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology* 16:179–189.
- Marin J., Battistuzzi F.U., Brown A.C., Hedges S.B. 2017. The Timetree of prokaryotes: new insights into their evolution and speciation. *Mol. Biol. Evol.* 34:437–446.
- Marshall C.R. 2019. Using the fossil record to evaluate Timetree timescales. *Front. Genet.* 10:1049.
- Martijn J., Vosseberg J., Guy L., Offre P., Ettema T.J.G. 2018. Deep phototrophic origin outside the sampled alphaproteobacteria. *Nature* 557:101–105.
- McCutcheon J.P., Moran N.A. 2012. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10:13–26.
- McGowen M.R., Tsagkogeorga G., Álvarez-Carretero S., dos Reis M., Struebig M., Deaville R., Jepson P.D., Jarman S., Polanowski A., Morin P.A., Rossiter S.J. 2020. Phylogenomic resolution of the cetacean tree of life using target sequence capture. *Syst. Biol.* 69:479–501.
- McLoughlin N., Wilson L.A., Brasier M.D. 2008. Growth of synthetic stromatolites and wrinkle structures in the absence of microbes - implications for the early fossil record. *Geobiology* 6:95–105.
- Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R., Teeling E. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37:1530–1534.
- Moody E.R., Mahendrarajah T.A., Dombrowski N., Clark J.W., Petitjean C., Offre P., Szöllösi 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:e66695.
- Muñoz-Gómez S.A., Susko E., Williamson K., Eme L., Slamovits C.H., Moreira D., López-García P., Roger A.J. 2022. Site-and-branch-heterogeneous analyses of an expanded data set favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* 6:253–262.
- Naser-Khdour S., Quang Minh B., Lanfear R. 2022. Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. *Syst. Biol.* 71:959–972.
- Nisbet E.G., Sleep N.H. 2001. The habitat and nature of early life. *Nature* 409:1083–1091.
- Pagel M., Meade A., Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* 53:673–684.
- 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. *Proc. Natl. Acad. Sci. U.S.A.* 108:13624–13629.
- Parks D.H., Rinke C., Chuvochina M., Chaumeil P.A., Woodcroft B.J., Evans P.N., Hugenholtz P., Tyson G.W. 2017. Recovery of nearly 8000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2:1533–1542.
- Pawitan Y. 2003. In *All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, UK: Oxford University Press.
- Perreau J., Moran N.A. 2022. Genetic innovations in animal-microbe symbioses. *Nat. Rev. Genet.* 23:23–39.
- Planavsky N.J., Asael D., Hofmann A., Reinhard C.T., Lalonde S.V., Knudsen A., Wang X., Ossa Ossa F., Pecoits E., Smith A.J.B., Beukes N.J., Bekker A., Johnson T.M., Konhauser K.O., Lyons T.W., Rouxel O.J. 2014. Evidence for oxygenic photosynthesis half a billion years before the great oxidation event. *Nat. Geosci.* 7:283–286.

- Quang L.S., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.
- Ravenhall M., Škunca N., Lassalle F., Dessimoz C. 2015. Inferring horizontal gene transfer. *PLoS Comput. Biol.* 11:e1004095.
- Revell L.J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12:e16505.
- Sánchez-Baracaldo P., Cardona T. 2020. On the origin of oxygenic photosynthesis and cyanobacteria. *New Phytol.* 225:1440–1446.
- Sánchez-Baracaldo P., Raven J.A., Pisani D., Knoll A.H. 2017. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl. Acad. Sci. U.S.A.* 114:E7737–E7745.
- Sánchez-Baracaldo P., Ridgwell A., Raven J.A. 2014. A neoproterozoic transition in the marine nitrogen cycle. *Curr. Biol.* 24:652–657.
- Schenk J.J. 2016. Consequences of secondary calibrations on divergence time estimates. *PLoS One* 11:e0148228.
- Schopf J.W. 2011. The paleobiological record of photosynthesis. *Photosynth. Res.* 107:87–101.
- Shih P.M., Hemp J., Ward L.M., Matzke N.J., Fischer W.W. 2017. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* 15:19–29.
- Strassert J.F.H., Irisarri I., Williams T.A., Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* 12:1879.
- Szöllösi G.J., Höhna S., Williams T.A., Schrempf D., Daubin V., Boussau B. 2022. Relative time constraints improve molecular dating. *Syst. Biol.* 71:797–809.
- Taib N., Megrian D., Witwinowski J., Adam P., Poppleton D., Borrel G., Beloin C., Gribaldo S. 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* 4:1661–1672.
- Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15:1647–1657.
- Toft C., Andersson S.G.E. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* 11:465–475.
- Tomitani A., Knoll A.H., Cavanaugh C.M., Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc. Natl. Acad. Sci. U.S.A.* 103:5442–5447.
- Wang S., Luo H. 2021. Dating Alphaproteobacteria evolution with eukaryotic fossils. *Nat. Commun.* 12:3324.
- Wang X., Planavsky N.J., Hofmann A., Saupe E.E., De Corte B.P., Philippot P., LaLonde S.V., Jemison N.E., Zou H., Ossa F.O., Rybacki K., Alfimova N., Larson M.J., Tsikos H., Fralick P.W., Johnson T.M., Knudsen A.C., Reinhard C.T., Konhauser K.O. 2018. A Mesoproterozoic shift in uranium isotope systematics. *Geochim. Cosmochim. Acta* 238:438–452.
- Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60:150–160.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.
- Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.* 14:717–724.
- Yang Z., Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 23:212–226.
- Yarza P., Yilmaz P., Priesse E., Glöckner F.O., Ludwig W., Schleifer K.H., Whitman W.B., Euzéby J., Amann R., Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* 12:635–645.
- Zhang H., Wang S., Liao T., Crowe S.A., Luo H. 2023. Emergence of *Prochlorococcus* in the Tonian oceans and the initiation of Neoproterozoic oxygenation. *bioRxiv*: 2023–2029.
- Zhang Y., Golubic S. 1987. Endolithic microfossils (Cyanophyta) from early Proterozoic stromatolites, Hebei, China. *Acta Micropalaeont. Sin.* 4:1–12.
- Zhu Q., Mai U., Pfeiffer W., Janssen S., Asnicar F., Sanders J.G., Belda-Ferre P., Al-Ghalith G.A., Kopylova E., McDonald D., Kosciolk T., Yin J.B., Huang S., Salam N., Jiao J.Y., Wu Z., Xu Z.Z., Cantrell K., Yang Y., Sayyari E., Rabiee M., Morton J.T., Podell S., Knights D., Li W.J., Huttenhower C., Segata N., Smarr L., Mirarab S., Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* 10:5477.
- Zuckerkandl E., Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: *Evolving genes and proteins*. New York, USA: Academic Press. p. 97–166.