

Log Transformation Improves Dating of Phylogenies

Uyen Mai,^{*,1} Siavash Mirarab²

¹Department of Computer Science and Engineering, UC San Diego, CA 92093, USA

²Department of Electrical and Computer Engineering, UC San Diego, CA 92093, USA

*Corresponding author: E-mail: smirarab@ucsd.edu

Associate Editor:

Abstract

Phylogenetic trees inferred from sequence data often have branch lengths measured in the expected number of substitutions and therefore, do not have divergence times estimated. These trees give an incomplete view of evolutionary histories since many applications of phylogenies require time trees. Many methods have been developed to convert the inferred branch lengths from substitution unit to time unit using calibration points, but none is universally accepted as they are challenged in both scalability and accuracy under complex models. Here, we introduce a new method that formulates dating as a non-convex optimization problem where the variance of log-transformed rate multipliers are minimized across the tree. On simulated and real data, we show that our method, wLogDate, is often more accurate than alternatives and is more robust to various model assumptions.

Key words: time tree, divergence time estimation, phylogenetic dating, molecular dating, non-convex optimization.

Introduction

Phylogenetic inference from sequence data does not reveal divergence time (i.e. exact timing of evolutionary events) unless paired with external timing information. Under standard models of sequence evolution, the evolutionary processes, including sequence divergence, are fully determined by the product of the absolute time and mutation rates in a non-identifiable form. Thus, these models measure branch lengths in the unit of expected numbers of

mutations per site (since standard models like GTR (Tavaré, 1986) only allow substitutions, focusing on these models, we use *substitutions* and *mutations* interchangeably throughout this paper). Nevertheless, knowing divergence times is crucial for understanding evolutionary processes (Forest, 2009; Hillis *et al.*, 1996) and is a fundamental need in many clinical applications of phylogenetics and phylodynamics (Volz *et al.*, 2013). A commonly used approach first infers a phylogeny with branch lengths in the unit of substitution per site and then dates the phylogeny by translating branch lengths from substitution

unit to time unit; co-estimation of topology and dates is also possible (Drummond *et al.*, 2006) though its merits have been debated (Wertheim *et al.*, 2010).

The fundamental challenge in dating is to find a way to factorize the number of substitutions into the product of the evolutionary rate and time. A common mechanism allowing this translation is to impose soft or hard constraints on the timing of *some* nodes of the tree, leaving the divergence times of the remaining nodes to be inferred based on the constrained nodes. Timing information is often in one of two forms: calibration points obtained from the geological record (Kodandaramaiah, 2011) and imposed on either internal nodes or tips that represent fossils (see Donoghue and Yang, 2016), or tip sampling times for fast-evolving viruses and bacteria. The constraints still leave us with a need to extrapolate from observed times for a few nodes to the remaining nodes, a challenging task that requires a mathematical approach. Obtaining accurate timing information and formulating the right method of extrapolation are both challenging (e.g., see Rutschmann, 2006).

Many computational methods for dating phylogenies are available (e.g., see Kumar and Hedges, 2016; Rutschmann, 2006), and a main point of differentiation between these methods is the clock model they assume (Sanderson, 1998). Some methods rely on a strict molecular clock (Zuckerkandl, 1962) where rates are effectively

assumed to be constant (e.g., Langley and Fitch, 1974; Shankarappa *et al.*, 1999). However, empirical evidence has now made it clear that rates can vary substantially, and ignoring these changes can lead to incorrect dating (Bromham and Penny, 2003; Kumar, 2005). Consequently, there have been many attempts to *relax* the molecular clock and allow variations in rates. A main challenge in relaxing the clock is the need for a model of rates, and it is not clear what model should be preferred. As a result, many methods for dating using relaxed molecular clocks have been developed. Some of these methods allow rates to be drawn independently from a stationary distribution (Akerborg *et al.*, 2008; Drummond *et al.*, 2006; Volz and Frost, 2017) while others model the evolution of rates with time (Huelsenbeck *et al.*, 2000) or allow correlated rates across branches (e.g., Drummond and Suchard, 2010; Kishino *et al.*, 2001; Lepage *et al.*, 2007; Sanderson, 2002; Snir *et al.*, 2012; Tamura *et al.*, 2012; Thorne *et al.*, 1998). Despite these developments, strict molecular clocks continue to be used, especially in the context of intraspecific evolution where there is an expectation of relatively uniform rates (Brown and Yang, 2011).

Another distinction between methods is the use of explicit models (Sanderson, 1997). Many dating methods use a parametric statistical model and formulate dating as estimating parameters in a maximum likelihood (ML) or Bayesian

inference framework (e.g., Drummond *et al.*, 2006; Langley and Fitch, 1974; To *et al.*, 2015; Volz and Frost, 2017). Another family of methods (e.g., Sanderson, 2002; Tamura *et al.*, 2012) formulate dating as optimization problems, including distance-based optimization (e.g., Xia, 2018; Xia and Yang, 2011), that avoid computing likelihood under an explicit statistical model. When the assumed parametric model is close to the reality, we expect parametric methods to perform well. However, these methods can be sensitive to model deviations, a problem that may be avoided by methods that avoid using specific models.

In this paper, we introduce LogDate, a new method of dating rooted phylogenies that allows variations in rates but without modeling rates using specific distributions. We define mutation rates necessary to compute time unit branch lengths as the product of a single global rate and a set of rate multipliers, one per branch. We seek to find the overall rate and all rate multipliers such that the log-transformed rate multipliers have the minimum variance. This formulation gives us a constrained optimization problem, which although not convex, can be solved in a scalable fashion using the standard approaches such as sequential least squares programming. While formulation of dating as an optimization problem is not new (Langley and Fitch, 1974; To *et al.*, 2015), here we introduce log-transformation of the rate multipliers, which as we will show,

results in more accurate dates. Our observation is in line with a recent change to RelTime (Tamura *et al.*, 2018) where the switch from arithmetic means to geometric means (between rates of sister lineages) has improved accuracy. In extensive simulation studies and three biological data sets, we show that a weighted version of LogDate, namely wLogDate, has higher accuracy in inferring node ages compared to alternative methods, including some that rely on time-consuming Bayesian inference. While wLogDate can date trees using both sampling times for leaves (e.g., in viral evolution) or estimated time of ancestors, most of our results are focused on cases with sampling times at the tips of the tree.

Methods

Definitions and notations

For a rooted binary tree T with n leaves, we give each node a unique index in $[0, \dots, 2n-2]$. By convention, the root is always assigned 0, the other internal nodes are arbitrarily assigned indices in the range $[1, \dots, n-1]$, and the leaves are arbitrarily assigned indices in the range $[n-1, \dots, 2n-2]$. In the rest of this paper, we will refer to any node by its index. If a node i is not the root node, we let $par(i)$ denote the parent of i and if i is not a leaf, we let $c_l(i)$ and $c_r(i)$ denote the left and right children of i , respectively. We refer to the edge connecting $par(i)$ and i as e_i .

We can measure each edge e_i of T in either time unit or substitution unit. Let t_i denote the divergence time of node i , i.e. the time when

species i diverged into $c_l(i)$ and $c_r(i)$. Then for any node i other than the root, $\tau_i = t_i - t_{par(i)}$ is the length of the edge e_i in time unit. We measure divergence time of a node with respect to a fixed reference point in the past (i.e., time increases forward). Thus, we enforce $t_i > t_{par(i)}$ for all i . Let μ_i be the substitution rate (per sequence site per time unit) on branch e_i , then the expected number of substitutions per sequence site is $b_i = \mu_i \tau_i$. Let $\tau = [\tau_1, \dots, \tau_{2n-2}]$ and $\mathbf{b} = [b_1, \dots, b_{2n-2}]$.

From sequence data, $\mathbf{b}$ can be inferred using standard methods such as maximum parsimony (Fitch, 1971), minimum evolution (Rzhetsky and Nei, 1993), neighbor-joining (Gascuel, 1997; Saitou and Nei, 1987), and maximum-likelihood (ML) (Felsenstein, 1981; Guindon *et al.*, 2010; Nguyen *et al.*, 2015). Note that inferred trees need to be rooted subsequently using an outgroup (that can be removed) or automatic methods such as midpoint or minimum variance rooting (Mai *et al.*, 2017). We let $\hat{b}_i$ denote the estimate of b_i by an inference method and let $\hat{\mathbf{b}} = [\hat{b}_1, \dots, \hat{b}_{2n-1}]$.

In this paper, we are interested in computing τ from $\hat{\mathbf{b}}$. The computation of τ from $\hat{\mathbf{b}}$ is complicated by two factors: (1) the possibility of change among rates, and (2) deviations of the inferred edge length $\hat{b}_i$ from the true value b_i .

To better describe the mathematical formulation of the optimization problem, we first do the following change of variables. Assuming the mutation rates on the branches are distributed around a global rate μ , we define

$\nu_i = \frac{\mu \tau_i}{b_i}$. Let $\mathbf{x} = [\nu_1, \dots, \nu_{2n-2}, \mu]$; our goal of finding τ is identical to finding $\mathbf{x}$.

Dating as a constrained optimization problem
We formulate dating as an optimization problem on $2n-1$ variables $\mathbf{x} = [\nu_1, \dots, \nu_{2n-2}, \mu]$, subject to the linear constraints defined by calibration points and/or sampling times. Many existing methods, including LF (Langley and Fitch, 1974) and LSD (To *et al.*, 2015), can be described in this framework, with the choice of the objective function distinguishing them from each other. We start by describing the setup of the constraints enforced by a set of calibration points/sampling times, and show that they can all be written as linear equations on $\mathbf{x}$. We then give the formulation of both LF and LSD in this framework and use their formulation to motivate our own new approach. Finally, we describe strategies to solve the wLogDate optimization problem.

Linear constraints Ψ from sampling times

For any pair of nodes (i, j) (where each of i and j can either be a leaf or an internal node) with enforced divergence times (t_i, t_j) , the following constraint $\psi(i, j)$ must be satisfied

$$\psi(i, j): \mu(t_j - t_i) = \sum_{k \in P(m, j)} \nu_k \hat{b}_k - \sum_{k \in P(i, m)} \nu_k \hat{b}_k \quad (1)$$

where m is the LCA of i and j and $P(m, j)$ and $P(i, m)$ are the paths connecting m to j and i to m , respectively. Thus, given k time points, $k(k-1)/2$ constraints must hold. However, only $k-1$ of these constraints imply all others, as we show below.

Let t_0 be the *unknown* divergence time at the root of the tree. For k calibration points $t_1, \dots, t_k$, we can setup k constraints of the form:

$$C_i: \mu(t_i - t_0) = \sum_{k \in P(0,i)} \nu_k \hat{b}_k, \quad (2)$$

where node 0 is the root and $P(0,i)$ is the path from the root to node i . For any pair (i,j) , the linear constraint given in Eq. 1 can be derived by subtracting C_i from C_j side by side. Also, we can remove t_0 from the set of constraints by subtracting C_1 from all other constraints $C_2, \dots, C_k$. This gives us the final $k-1$ linear constraints on $\mathbf{x}$, which we denote as Ψ . We can build Ψ using Algorithm 1 (Supplementary material).

Optimization Criteria

Since $\nu_i = \frac{\mu\tau_i}{\hat{b}_i}$, the distribution of ν_i is influenced by both the distribution of the rates (μ_i) and the distribution of $\hat{b}_i$ around b_i . In traditional strict-clock models (Zuckerkandl, 1962), a constant rate is assumed throughout the tree ($\forall_i \mu_i = \mu$). Under this model, the distribution of ν_i is determined by deviations of $\hat{b}_i$ from b_i .

Langley and Fitch (1974) (LF) modeled the number of *observed* substitutions *per sequence site* on a branch i by a Poisson distribution with mean $\lambda = \mu\tau_i$ and treated $s\hat{b}_i$ as if they were the total number of *observed* substitutions; as such, they assume $s\hat{b}_i \sim \text{Poisson}(s\mu\tau_i)$, where s is the sequence length. Therefore, by changing variable,

we can write the log-likelihood function as:

$$\sum_{i=1}^{2n-2} \left(s\hat{b}_i \log(s\hat{b}_i) - \log((s\hat{b}_i)!) \right) + \sum_{i=1}^{2n-2} s\hat{b}_i (\log \nu_i - \nu_i).$$

Given s and $\hat{b}_i$, LF finds $\mathbf{x}$ that maximizes the log-likelihood function and subject to the constraints Ψ . As such,

$$\mathbf{x}_P^* = \arg \min_{\mathbf{x}} \sum_{i=1}^{2n-2} \hat{b}_i (\nu_i - \log \nu_i) \text{ subject to } \Psi. \quad (3)$$

To *et al.* (2015) assume $\hat{b}_i$ follows a Gaussian model: $\hat{b}_i \sim \text{Gaussian}(\mu\tau_i, \sigma_i^2)$ and assume the variance is approximated by $\frac{\hat{b}_i}{s}$ (the method includes smoothing strategies omitted here). Then, the negative log likelihood function can be written as:

$$\sum_{i=1}^{2n-2} \frac{(\hat{b}_i - \mu\tau_i)^2}{\sigma_i^2} \approx \sum_{i=1}^{2n-2} \frac{s}{\hat{b}_i} (\hat{b}_i - \mu\tau_i)^2 = \sum_{i=1}^{2n-2} s\hat{b}_i (1 - \nu_i)^2.$$

Thus, the ML estimate can be formulated as:

$$\mathbf{x}_G^* = \arg \min_{\mathbf{x}} \sum_{i=1}^{2n-2} \hat{b}_i (1 - \nu_i)^2 \text{ subject to } \Psi. \quad (4)$$

Both LF and LSD have convex formulations. Langley and Fitch (1974) proved that their negative log-likelihood function is convex and thus the local minimum is also the global minimum. Our constraint-based formulation of LF also can be easily proved convex by showing its Hessian matrix is positive definite. To *et al.* (2015) pointed out their objective function is a weighted least squares. Using our formulation, we also see that Eq. 4 together with the calibration constraints form a standard convex quadratic optimization problem which has a unique analytical solution.

LogDate Method

Motivation

LF only seeks to model the errors in $\hat{\mathbf{b}}$ and ignore true rate heterogeneity. Strict-clock assumption is now believed to be unrealistic in many settings (Ho, 2014; Pulquério and Nichols, 2007; Schwartz and Maresca, 2006), motivating relaxed clocks, typically by assuming that μ_i s are drawn i.i.d. from some distribution (e.g., Akerborg *et al.*, 2008; Drummond *et al.*, 2006; Volz and Frost, 2017). Most methods rely on presumed parametric distributions (typically, LogNormal, Exponential, or Gamma) and estimate parameters using ML (Volz and Frost, 2017), MAP (Akerborg *et al.*, 2008), or MCMC (Drummond and Rambaut, 2007; Drummond *et al.*, 2006). The LSD method, which like LF directly models errors in $\hat{\mathbf{b}}$, is additionally justified under a normally-distributed clock model. Choices of specific distributions in these methods are not motivated by the knowledge that real data follow them exactly (for example, the Normal distribution has to be misspecified as mutation rates cannot be negative).

Our goal is to avoid explicit parameter inference under a model of rate multipliers. Instead, we follow the assumption shared by existing methods like LSD and LF: we assume that given two solutions of $\mathbf{x}$ both satisfying the calibration constraints, the solution with less variability in ν_i values is preferable. Thus, we prefer solutions that minimize deviations from a strict clock while allowing deviations. A natural way to minimize

deviations from the clock is to minimize the variance of $\frac{\tau_i}{b_i}$. This can be achieved by finding μ and all ν_i such that ν_i is centered at 1 and $\sum_{i=1}^{2n-2}(\nu_i - 1)^2$ is minimized. Interestingly, the ML function used by LSD (Eq. 4) is a weighted version of this approach.

The minimum variance principle results in a fundamental asymmetry: multiplying or dividing the rate of a branch by the same factor are penalized differently (Fig 1a). For example, the penalty for $\nu_i=4$ is more than ten times larger than $\nu_i=1/4$. The LF model is more symmetrical than LSD but remains asymmetrical (Fig 1a). This asymmetry results from the asymmetric distribution of the Poisson distribution around its mean, especially for small mean, in log scale (Fig 1b). Because of this asymmetry, methods like LSD and LF judge a very small $\hat{b}_i/b_i$ to be within the realm of possible outcomes, and thus penalize $\nu_i < 1$ multipliers less heavily than $\nu_i > 1$.

Our method is based on a principle, which we call the *symmetry of ratios*: the penalty for multiplying a branch by a factor of ν should be no different than dividing the branch by ν . Note that this assertion is only applicable to true variations of the mutation rate (i.e, ignoring branch length estimation error). We further motivate this principle with more probabilistic arguments below, but here we make the following case. If one considers the distribution of rate *multipliers* for various branches, absent of an explicit model, it is reasonable to assume that

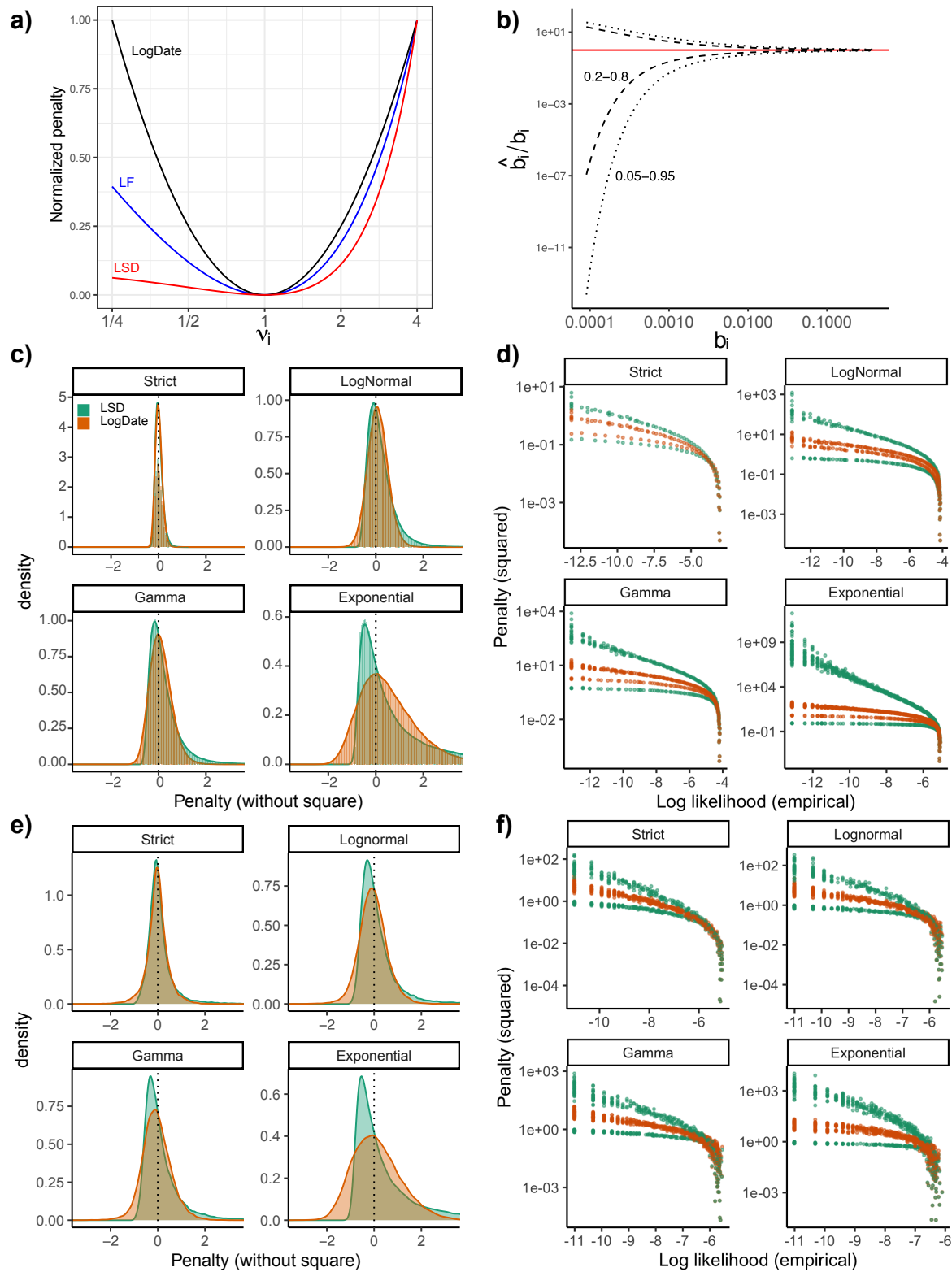


FIG. 1. (a) The penalty associated to multiplying a single edge i with multiplier ν_i in LSD, LF, and LogDate approaches, as shown in Equations 3, 4, and 5. To allow comparison, we normalize the penalty to be zero at $\nu=1$ and to be 1 at $\nu=4$. (b) The confidence interval of the ratio between estimated and true branch length using the Poisson model. For this purpose of this exposition, we assume that the estimated branch length equals the number of substitutions occurring on the branch and follows a Poisson distribution (i.e., JC69 model), divided by sequence length. With these assumptions, the CI for estimate

length $\hat{b}_i$ is between $1/2\chi^2_{2sb_i}$ and $1/2\chi^2_{2sb_i+2}$; we draw the CI for $\alpha/2=0.05$ and $\alpha/2=0.2$ to get 0.2–0.8 and 0.05–0.095 intervals

for $0.0001 \leq b_i \leq 0.4$. (c) and (e): Density and histograms of penalty terms (without square) used by LSD ($\mu\tau_i/b_i - 1$) and LogDate ($\log(\mu\tau_i/b_i)$) under different clock models. (c) Fixing $\mu\tau_i=0.1$, we draw 500000 rate multipliers (r_i) from LogNormal, Gamma, or Exponential distributions with mean 1 and variance 0.16 for LogNormal and Gamma. For strict clock, $r_i=1$. We then draw estimated branch length for each replicate i from the Normal distributed with mean $b_i=r_i\mu\tau_i$ and variance b_i/s for $s=200$. (e) The branch lengths are estimated from the sequences using PhyML from simulated sequences of (TQ *et al.*, 2015), as explained in the text. Parameters of rate multiplier distributions match part (c). We omit extremely short

compared to an overall rate, branches rates are as likely to increase by a factor of ν as they are to decrease by a factor of ν . When this statement is true, we shall prefer a method that penalizes ν and $1/\nu$ identically. To ensure the symmetry of ratios, we propose taking the logarithm of the multipliers ν_i before minimizing their variance. Minimizing the variance of the rates in log-scale is the essence of our method. It achieves the symmetry, and, as we show below, a better correspondence between penalty and data likelihood.

Log-transformation has long been used to reduce data skewness before applying linear regression (Clifford *et al.*, 2013; Keene, 1995; Stynes *et al.*, 1986; Xiao *et al.*, 2011). In molecular dating, it can be argued that log-transformation is implicit in the new version of RelTime (Tamura *et al.*, 2018) where the geometric means between sister lineages replaced the arithmetic means in its predecessor. The improvement in the accuracy of RelTime encourages a wider use of log-transformation in molecular dating. Note that log-transforming the rate multipliers before minimizing their least squares penalty is identical to applying linear least squares after log-transformation of both time and the number of substitutions. In other fields, log-transformation has been used to make the least-squares method more robust to highly skewed distributions (Aban and Meerschaert, 2004; Meaney *et al.*, 2007).

LogDate optimization function

We formulate the LogDate problem as follows. Given $\hat{\mathbf{b}}$ and the set of calibration constraints described earlier, we seek to find

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{2n-1} \log^2(\nu_i) \quad \text{subject to } \Psi. \quad (5)$$

This objective function satisfies the symmetry of ratio property (Fig. 1a). Since ν_i values are multipliers of rates around μ , if we assume μ is the mean rate, the LogDate problem is equivalent to minimizing the variance of the log-transformed rate multipliers (around their mean 1). The objective function only depends on ν_i ; however, note that μ is still included in the constraints and therefore is part of the optimization problem. This setting reduces the complexity of the objective function and speeds up the numerical search for the optimal solution. Since the values of ν_i close to 1 are preferred in Eq. 5, the optimal solution would push μ to the mean rate.

Justification as a relaxed-clock model.

After log-transformation, LogDate, similar to LSD, constructs the objective function using the least squares principle (for ease of exposition, here we discuss ordinary least-squares without weights). We can rewrite the objective function of LSD as $\sum_i (\frac{\mu\tau_i}{b_i} - 1)^2$ and that of LogDate as $\sum_i (\log \frac{\mu\tau_i}{b_i})^2$ and see that both seek to find a global rate μ and the time τ_i for each branch to minimize the total deviations of the estimated branches from $\mu\tau_i$. This observation may motivate viewing both LSD and LogDate as strict-clock

methods. However, the following result justifies viewing LogDate as a relaxed clock method.

We can prove that if the mutation rates μ_i are drawn i.i.d. from a LogNormal distribution with any parameters with mode μ and the branches are estimated without error (i.e. $\hat{b}_i = b_i$ for all i), then ν_i follows a LogNormal distribution with mode 1 and the LogDate optimization problem is equivalent to finding ν that have maximum joint probability, subject to the constraints. The proof is given in Claim 1 (Supplementary materials).

Justification for symmetry of ratios.

Having shown that LogDate has a justification under the LogNormal distribution, we now compare LogDate and LSD objective functions in a wider range of clock models. Recall that the objective functions of LSD and LogDate are the sum-of-squares of their penalty terms, which are $\frac{\mu\tau_i}{b_i} - 1$ for LSD and $\log \frac{\mu\tau_i}{b_i}$ for LogDate.

Following the likelihood principle, an ideal objective function must assign equal penalties to data values that are equally likely to occur. Therefore, for an ideal objective function written as sum-of-squares of the penalty terms, the probability distribution of its penalty terms (before square) under the model that generates the data must be symmetric around 0 (because of the square). The true distribution of our penalty terms is a function of both clock rate variations and branch length estimation error. While no objective function is ideal for all

compound models of rates and estimation error, a robust objective function should remain close to symmetric and maintain a low skewness under a wide range of models. We now present several theoretical and empirical results comparing LogDate and LSD in terms of skewness of distributions of their penalty terms.

First, consider a relaxed clock model of the rates and assume no branch estimation error (i.e., $\hat{b}_i = \mu_i \tau_i$). If μ_i follows a LogNormal distribution parameterized by θ and σ then it is easy to see that $\frac{\mu\tau_i}{b_i} = \frac{\mu}{\mu_i}$ (penalty of LSD) also follow a LogNormal distribution and the skewness depends on σ . In contrast, the $\log \frac{\mu\tau_i}{b_i}$ (penalty of LogDate) follows a Normal distribution, which has skewness 0, and for which least square estimation is the maximum likelihood estimator. Thus, as stated before, log-transforming is the ML solution if rate multipliers are log-normally distributed.

Now assume μ_i follows a Gamma distribution with mean μ . Then $\mu\tau_i/\hat{b}_i = \mu/\mu_i$ follows an Inverse Gamma distribution while its log-transformation follows a Log-Gamma distribution. We can analytically compute the skewness of the penalty terms of LSD and LogDate and compare them (see Supplementary materials for the equations). As shown in Fig. S1 (Supplementary materials), the skewness of LSD is much higher than that of LogDate, especially for higher variance of the gamma rates. Higher skewness of penalty terms violates the likelihood principle mentioned before. Thus, for the two models where we could compute

analytical formulas for skewness, we have grounds to prefer LogDate.

Next, we consider the compound impacts of branch length estimation error and rate variation, and we study the question in two ways. One approach is to measure the combined effect of error and true variation by simulating sequence data and measuring $\hat{b}_i$ for known b_i empirically; here, we use simulations by To *et al.* (2015) with 1000 sites and PhyML-inferred trees (details will be provided in the Experiments section). The other approach is modelling the compound effect. While it is hard to know generally how estimated branch length is distributed around its expected value, here, we can follow To *et al.* (2015) and assume $\hat{b}_i \sim \mathcal{N}(b_i, \frac{b_i}{s})$. The other challenge is that the compound distribution of estimation error and rate multipliers is hard to compute analytically. However, we can easily generate a very large number of samples from compound distributions and analyze the empirical distribution to approximate the true distribution.

Inspecting the empirical density of the penalty terms of LSD and LogDate across different clock models result in consistent patterns using both approaches, modeling the compound distribution (Fig. 1c) and using simulated sequence data (Fig. 1e). Across three models of rates, Exponential, LogNormal, and Gamma, the distributions of the LogDate penalty terms are always more symmetric than that of LSD. Results are similar for other rate models such

as Log-Uniform and are further amplified when the variance is increased (Fig. S2a, Supplementary Materials).

To further explore that relationship between the likelihood and the penalty assigned by LogDate and LSD, we plot the penalty (with square terms) versus the empirical log likelihood of the rate multipliers (Figs. 1d and 1f and S2b in Supplementary Materials). Ideally, increasing likelihood should monotonically decrease penalty, and points with similar likelihood should have similar penalties. In both modelled and simulated branch lengths and across models, LSD assigns two sets of widely different penalties (one for increased and one for decreased rates) to data with similar likelihood. LogDate, while far from perfect, is much closer to the ideal mapping between likelihood and penalty. Also, for LogNormal with *median* rate multipliers set to 1, we empirically observe a perfectly monotonic relationship between the penalty and likelihood (Fig. S2b), as theory suggested.

wLogDate optimization function

The simple LogDate formulation, however, has a limitation: by allowing rates to vary freely in a multiplicative way, it fails to deal with the varied levels of relative branch error; i.e., the ratio of the estimated branch length to the true branch length ($\hat{b}_i/b_i$). As $\hat{b}_i$ is estimated from the sequences, the error of $\hat{b}_i$ is directly related to the variations in the number of substitutions occurred along the

branch b_i . Let us assume sequences follow the Jukes and Cantor (1969) model, and let N_i be the total number of substitutions occurred along branch i on a sequence with length s . Under Jukes-Cantor model, we have $N_i \sim \text{Poisson}(s\mu\tau_i)$ and therefore, $\text{var}(N_i) = s\mu\tau_i$. Therefore, the variance of the *expected number of substitutions* around the true branch length is $\text{var}(\frac{N_i}{sb_i}) = \frac{s\mu\tau_i}{s^2b_i^2} = \frac{1}{b_i s}$. As Figure 1b shows, when b_i is small, $\frac{N_i}{s}$ can easily vary by several orders of magnitude around b_i . Furthermore, the distribution is not symmetric: drawing values several factors smaller than the mean is more likely than drawing values above the mean by the same factor. These analyses predict that the distribution of $\frac{\hat{b}_i}{b_i}$ depends strongly on b_i - with smaller b_i giving higher variance - and is not symmetric.

The variances of the relative error $\frac{\hat{b}_i}{b_i}$ is difficult to compute analytically due to the involvement of the sequence substitution model and the method to estimate $\hat{b}_i$, which are both unknown. Therefore, we instead use empirical analyses of the estimated branch lengths by PhyML to demonstrate our arguments. Consistent with our prediction, Figures S7 a and c illustrate that the relative error $\frac{\hat{b}_i}{b_i}$ varies more in small branches and the distribution is not symmetric. These properties of the branch length estimates are not modeled in our LogDate formulation and we seek to incorporate them in a refined version of LogDate which will be described below.

Since the true branch length b_i is unknown, a common practice is to use the estimated $\hat{b}_i$ in place of b_i to estimate its variance as $\frac{1}{\hat{b}_i s}$. This explains why both LF and LSD objective functions (Eqs. 3 and 4) have a weight of $\hat{b}_i$ for each term of ν_i . Following the same strategy, we propose weighting each $\log^2(\nu_i)$ term in a way that reduces the contribution of short branches to the total penalty, and thus allows more deviations in the log space if the branch is small (and is thus subject to higher error). Since we log-transform ν_i and pursue a model-free approach, explicitly computing the weights to cancel out the variations of relative error among the branches is challenging. However, since the weights should reflect the variance of $\frac{\hat{b}_i}{b_i}$ (logarithmic scale), they should monotonically increase with $\hat{b}_i$ (Fig. 1b) to allow more variance for the *relative* errors in short branches than in long branches. We use $\sqrt{\hat{b}_i}$ as weights, a selection driven by simplicity and empirical performance (shown in a later section).

The shortest branches require even more care. When the branch is very short, for a limited-size alignment, the evolution produces zero mutations with high probability. For these no-event branches, tree estimation tools report arbitrary small lengths (see Fig. S7), rendering $\hat{b}_i$ values meaningless for very small branches. To deal with this challenge, the r8s's implementation of LF (Sanderson, 2003) collapses all branches with length $\hat{b}_i < 1/s$. To *et al.* (2015) proposed adding a smoothing constant c/s to each $\hat{b}_i$

to estimate the variance of $\hat{b}_i$, where c is a parameter that the user can tune. Following a similar strategy, we propose adding a small constant $\tilde{b}$ to each $\hat{b}_i$. We choose $\tilde{b}$ to be the maximum branch length that produces no substitutions with probability at least $1-\alpha$ for $\alpha \in [0,1]$. Recall that N is the total number of *actual* substitutions on a branch. Under the Jukes and Cantor (1969) model, it is easy to show that $\arg\max_b Pr(N=0|b=\tilde{b}) \geq 1-\alpha = -\frac{1}{4}\log(1-\alpha)$. We choose this value as $\tilde{b}$ and set $\alpha=0.01$ by default. Thus, we define the weighted LogDate (wLogDate) as follows:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \sum_{i=1}^{2n-1} \sqrt{\tilde{b} + \hat{b}_i} \log^2(\nu_i) \quad (6)$$

subject to Ψ .

Solving the optimization problem

Both LogDate and wLogDate problems (Eq. 5 and Eq. 6) are non-convex, and hence solving them is non-trivial. The problem is convex if $0 \leq \nu_i \leq e$. For small clock deviation and small estimation error in $\hat{b}_i$, the ν_i values should be small so that the problem becomes convex with one local minimum. However, as $\nu_i \leq e$ is not guaranteed, we have to rely on gradient-based numerical methods to search for multiple local minima and select the best solution we can find. To search for local minima, we use the Scipy solver with **trust-constr** (Lalee *et al.*, 1998) method. To help the solver work efficiently, we incorporate three techniques that we next describe.

Computing Jacobian and Hessian matrices analytically helps speedup the search. By taking the partial derivative of each ν_i , we can compute the Jacobian, J , of Eq. 6. Also, since Eq. 6 is separable, its Hessian H is a $(2n-2) \times (2n-2)$ diagonal matrix. Simple derivations give us:

$$J = \left[2\sqrt{\tilde{b} + \hat{b}_1} \frac{\log \nu_1}{\nu_1}, \dots, 2\sqrt{\tilde{b} + \hat{b}_{2n-2}} \frac{\log \nu_{2n-2}}{\nu_{2n-2}} \right]^T$$

and $H_{ii} = 2\sqrt{\tilde{b} + \hat{b}_i} \frac{1 - \log \nu_i}{\nu_i^2}$.

Sparse matrix representation further saves space and computational time. The Hessian matrix is diagonal, allowing us to store only the diagonal elements. In addition, the constraint matrix defined by Ψ is highly sparse. If all sampling times are given at the leaves, the number of non-zero elements in our $(n-1) \times (2n-1)$ matrix is $O(n \log n)$ (Claim 3; Supplementary materials). If the tree is either caterpillar or balanced, the number of non-zeroes reduced to $\Theta(n)$. Thus, we use sparse matrix representation implemented in the Scipy package. This significantly reduces the running time of LogDate.

Starting from multiple feasible initial points is necessary given that our optimization problem is non-convex. Providing initial points that are feasible (i.e. satisfied the calibration constraints) helps the SciPy solver work efficiently. We designed a heuristic strategy to find multiple initial points given sampling times $t_1, \dots, t_n$ of all the leaves (as is common in phylodynamics).

We first describe the process to get a single initial point. We compute the root age t_0 and μ using root-to-tip regression (RTT) (Shankarappa *et al.*, 1999). Next, we scale all branches of T to conform with Ψ as follow: let $m = \operatorname{argmin}_i t_i$ (breaking ties arbitrarily). Let $d(r, i)$ denote the distance from the root r to node i and $P(r, m)$ denote the path from r to m . For each node i in $P(r, m)$, we set $\tau_i = \hat{b}_i(t_m - t_0)/d(r, m)$. Then going upward from m to r following $P(m, r)$, for each edge (i, j) we compute $t_j = t_i - \tau_i$ and recursively apply the process on the clade i . At the root, we set t_m to the second oldest (second minimum) sampling time and repeat the process on a new path until all leaves are processed. Since RTT gives us μ , to find ν we simply set $\nu_i = \mu\tau_i/\hat{b}_i$.

To find multiple initial points, we repeatedly apply RTT to a set of randomly selected clades of T and scale each clade using the aforementioned strategy. Specifically, we randomly select a set S of some internal nodes in the tree and add the root to S . Then, by a post-order traversal, we visit each $u \in S$ and date the clade u using the scaling strategy described above. We then remove the entire clade u from the tree but keep the node u as a leaf (note that the age of u is already computed) and repeat the process for the next node in S . The root will be the last node to be visited. After visiting the root, we have all the τ_i for all i . After having all the branches in time unit, we find $\mathbf{x}$ to minimize either Eq. 5 or Eq. 6, depending on whether LogDate or wLogDate is

chosen. In a tree of n leaves, we have $2^{(n-1)} - 1$ ways to select the initial non-empty set S , giving us enough room for randomization.

Computing confidence interval

With the ability of wLogDate to work on any combination of sampling times/calibration points on both leaves and internal nodes (as long as at least two time points are provided), we design a method to estimate the confidence intervals for the estimates of wLogDate. We subsample the sampling times/calibration points given to us repeatedly to create N replicate datasets (where N is 100 by default, but can be adjusted). Note our subsampling is not exactly a bootstrapping procedure as node sampling times cannot be resampled with replacement. We then compute the time tree for each replicate to obtain N different estimates for the divergence time of each node, from which we can compute their confidence intervals (95% as default). This sampling would work best when we have a fairly large number of calibration points, which is the case in phylodynamic settings where all (or nearly all) sampling times for the leaves are given, or in large phylogenies where abundant calibration points can be obtained from fossils. Although we refer to the resulting intervals as confidence intervals, it is important to recognize that the resampling procedure is not strictly justified via bootstrap theory because subsampling is

necessarily without replacement and sampled nodes are not independent of each other.

Experiments on simulated data

Phylogenetics setting

To *et al.* (2015) simulated a dataset of HIV *env* gene. Their time trees were generated based on a birth-death model with periodic sampling times. There are four tree models, namely D995_11_10 (M1), D995_3_25 (M2), D750_11_10 (M3), and D750_3_25 (M4), each of which has 100 replicates for a total of 400 different tree topologies. M1 and M2 simulate intra-host HIV evolution and are ladder-like while M3 and M4 simulate inter-host evolution and are balanced. Also, M4 has much higher root-to-tip distance (mean: 57) compared to M1–M3 (22, 33, and 29). Starting from conditions simulated by To *et al.* (2015), we use the provided time tree to simulate the clock deviations. Using an uncorrelated model of the rates, we draw each rate from one of three different distributions, each of which is centered at the value $\mu=0.006$ as in To *et al.* (2015). Thus, we set each μ_i to $x_i\mu$ where x_i is drawn from one of three distributions: LogNormal (mean:1.0, std: 0.4), Gamma ($\alpha=\beta=6.05$), and Exponential ($\lambda=1$). Sequences of length 1000 were simulated for each of the model conditions using SeqGen (Rambaut and Grass, 1997) under the same settings as To *et al.* (2015).

Calibrations on autocorrelated rate model.

We used the software NELSI and the same protocol as in (Ho *et al.*, 2015) to simulate a

dataset where the rates are autocorrelated. The dataset has 10 replicates, each contains 50 taxa. The time trees were generated under Birth-death model and the rate heterogeneity through time is modeled by the autocorrelation model (Kishino *et al.*, 2001)) with the initial rate set to 0.01 and the autocorrelated parameter set to 0.3. DNA sequences (1000 bases) were generated under Jukes-Cantor model. We used PhyML (Guindon *et al.*, 2010) to estimate the branch lengths in substitution unit from the simulated sequences while keeping the true topology. These trees are the inputs to wLogDate, RelTime, LF, and DAMBE (Xia, 2018) to infer time trees.

Real biological data

H1N1 2009 pandemic. We re-analyze the H1N1 biological data provided by To *et al.* (2015) which includes 892 H1N1pdm09 sequences collected worldwide between 13 March 2009 and 9 June 2011. We reuse the estimated PhyML (Guindon *et al.*, 2010) trees, 100 bootstrap replicates, and all the results of the dating methods other than LogDate that are provided by To *et al.* (2015).

San Diego HIV. We study a dataset of 926 HIV-1 subtype B *pol* sequences obtained in San Diego between 1996 and 2018 as part of the PIRC study. We use IQTree (Nguyen *et al.*, 2015) to infer a tree under the GTR+ Γ model, root the tree on 22 outgroups, then remove the outgroups. Because of the size, we could not run BEAST.

West African Ebola epidemic. We study the dataset of Zaire Ebola virus from Africa, which includes 1,610 near-full length genomes sampled between 17 March 2014 and 24 October 2015. The data was collected and analyzed by Dudas *et al.* (2017) using BEAST and re-analyzed by Volz and Frost (2017) using IQTree to estimate the ML tree and *treedater* to infer node ages. We run LSD, LF, and wLogDate on the IQTree from Volz and Frost (2017) and use the BEAST trees from Dudas *et al.* (2017), which include 1000 sampled trees (BEAST-1000) and the Maximum clade credibility tree (BEAST-MCC). To root the IQTree, we search for the rooting position that minimizes the triplet distance (Sand *et al.*, 2013) between the IQTree and the BEAST-MCC tree.

Methods Compared

For the phylodynamics data, we compared wLogDate to three other methods: LSD (To *et al.*, 2015), LF (Langley and Fitch, 1974), and BEAST (Drummond and Rambaut, 2007). For all methods, we fixed the true rooted tree topology and only inferred branch lengths. For LSD, LF, and wLogDate, we used phyML (Guindon *et al.*, 2010) to estimate the branch lengths in substitution unit from sequence alignments and used each of them to infer the time tree. LSD was run in the same settings as the QPD* mode described in the original paper (To *et al.*, 2015). LF was run using the implementation in r8s (Sanderson, 2003). wLogDate was run with

10 feasible starting points. For the Bayesian method BEAST, we also fixed the true rooted tree topology and only inferred node ages. Following To *et al.* (2015), we ran BEAST using HKY+ Γ 8 and coalescent with constant population size tree prior. We used two clock models on the rate parameter: the strict-clock (i.e. fixed rate) model and the LogNormal model. For the strict-clock prior, we set clock rate prior to a uniform distribution between 0 and 1. For the LogNormal prior, we set the ucl.d.mean prior to a uniform distribution between 0 and 1, and ucl.d.stdev prior to an exponential distribution with parameter $1/3$ (default). We always set the length of the MCMC chain to 10^7 generations, burn-in to 10%, and sampling to every 10^4 generations (identical to To *et al.* (2015)).

For the autocorrelated rate model, we compared wLogDate to LF and RelTime (Tamura *et al.*, 2018), which is one of the state-of-the-art model-free dating methods. We randomly chose subsets of the internal nodes (10% on average) as calibration points and created 20 tests for each of the 10 replicates (for a total of 200 tests). We also compared wLogDate to DAMBE using this dataset. Because DAMBE can only be run in interactive mode where each calibration point has to be manually placed onto the tree, we could not run DAMBE on the 200 tests with hundreds of calibration points in total. Therefore, we instead ran DAMBE only once on each of the 10 trees and infer a unit time tree for each of them (i.e.

calibrate the root to be at 1 unit time backward) and compared the results to that of wLogDate. DAMBE does not accept identical sequences so we removed identical sequences from the simulated alignments and trees before running DAMBE and ran wLogDate using these reduced trees to have a fair comparison.

Evaluation Criteria

On the simulated phylodynamics dataset where the ground truth is known, we compare the *accuracy* of the methods using several metrics. We compute the root-mean-square error (RMSE) of the true and estimated vector of the divergence times (τ) and normalize it by tree height. We also rank methods by RMSE rounded to two decimal digits (to avoid different ranks when errors are similar). In addition, we examine the inferred divergence time of the Most Recent Common Ancestor (tMRCA) and mutation rate. The comparison of methods mostly focuses on point-estimates of these parameters and the accuracy of the estimates (as opposed to their variance). In one analysis, we also compare the confidence intervals produced by wLogDate and BEAST on one model condition (M3 with LogNormal rate distribution). Finally, we examine the correlation between variance of the error in wLogDate and divergence times and branch lengths.

On the simulated data with autocorrelated rate, we show the distributions of the divergence times estimated by wLogDate, LF, and RelTime and

report the RMSE normalized by tree height for each replicate. To compare to DAMBE in inferring unit time trees, we report the average relative error of the inferred to the true divergence times. After removing identical sequences, there are 438 internal nodes in total across the 10 tree replicates. For each internal nodes, we compute the relative error of its divergence time inferred by either DAMBE or wLogDate to its true divergence time in the normalized true time tree, which is $\frac{|\hat{t}_i - t_i|}{t_i}$ where $\hat{t}_i$ and t_i are the inferred and true divergence times of node i , respectively. We report the average relative error per tree replicate and the average of all 438 nodes for DAMBE and wLogDate.

On real data, we show lineage-through-time (LTT) plots (Nee *et al.*, 1994), which trace the number of lineages at any point in time and compare tMRCA times to the values reported in the literature. We also compare the runtime of wLogDate to all other methods in all analyses.

Results

Simulated data for phylodynamics

We first evaluate the convergence of the ScipPy solver across 10 starting points (Fig. S3a). LogDate and wLogDate converge to a stable result after 50–200 iterations, depending on the model condition. Convergence seems easier when rates are Gamma or LogDate and harder when the rates are Exponential. Next, to control for the effect of the starting points on the accuracy of our method, we compare the error of these starting points

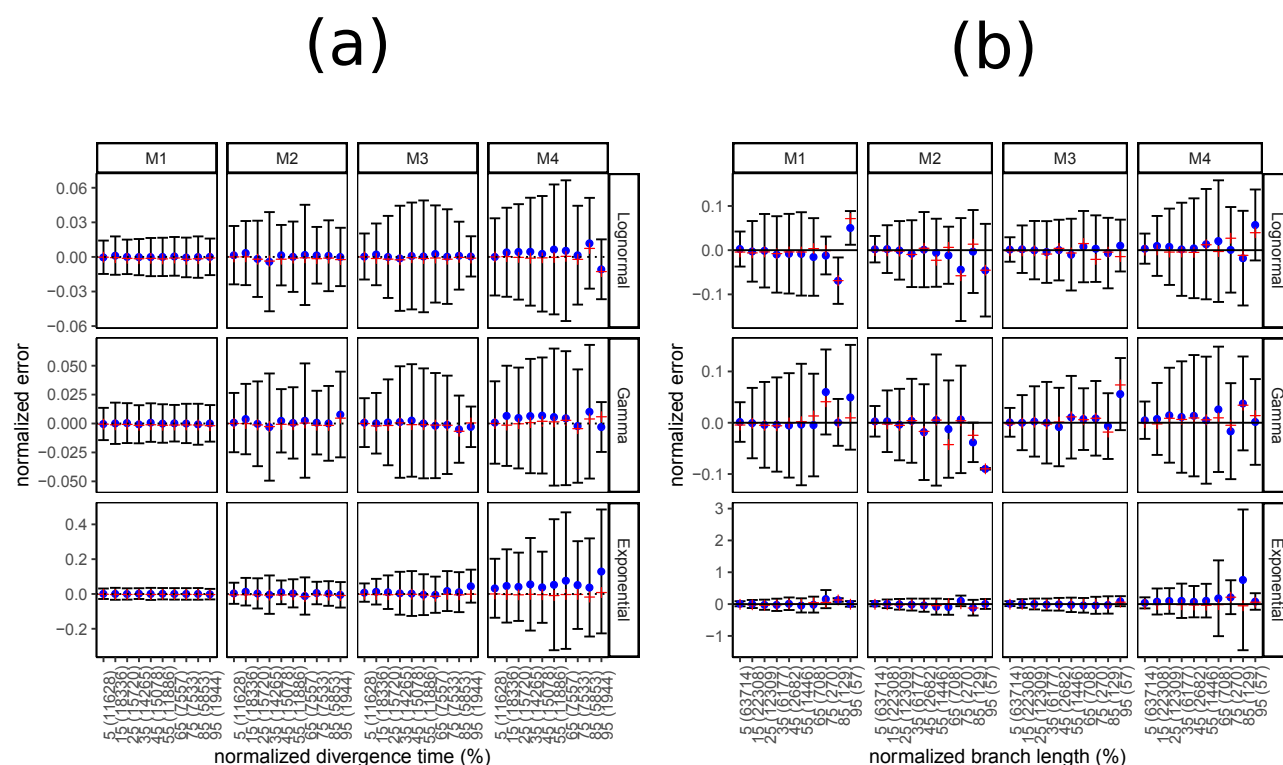


FIG. 2. Analyses of wLogDate on inferring branch lengths on simulated data. (a) error normalized by tree height versus divergence time (i.e. the time of the midpoint of each branch); both axes are normalized by the tree height. (b) error versus branch length (in time unit); both axes are normalized by the maximum branch length. For both (a) and (b), the x-axis is discretized into 10 bins of equal size. We label the bins by their median values, relative to either the tree height for (a) or the maximum branch length for (b). We also show the number of points in each bin in parentheses. Note the small number of points in the final bins in panel (b). For each bin, the blue dot represents the mean, the red cross represents the median, and the bar represents one standard deviations around the mean.

with the wLogDate optimal point (Fig. S3b). In all model conditions, the optimal point shows dramatic improvement in accuracy compared to the starting point. We then compare different weighting strategies for LogDate (Table S4). In all model conditions, the weighting $\sqrt{\hat{b}_i + \tilde{b}}$ is one of the two best, so it is chosen as the default weighting for wLogDate. Moreover, wLogDate is never worse than LogDate, and under exponential clock models, appropriate weighting results in dramatic improvements (Table S4).

model	Clock model	B.lnorm	B.strict	LF	LSD	wLogDate
M4	LogNormal	1	3	4	5	1
	Gamma	2	4	3	5	1
	Exponential	4	3	2	5	1
M3	LogNormal	2	3	3	5	1
	Gamma	4	2	2	5	1
	Exponential	5	3	2	4	1
M2	LogNormal	5	1	3	4	2
	Gamma	4	1	3	5	2
	Exponential	4	1	2	5	3
M1	LogNormal	4	1	2	4	2
	Gamma	5	1	1	4	1
	Exponential	2	1	3	3	5
average rank		3.5	2	2.5	4.5	1.75

Table 1. Ranking of the dating methods under different model conditions. For each model condition, the average RMSE of all internal node ages is computed and ranked among the dating methods (rounded to two decimal digits). The best method is shown in bold.

Next, we study the properties of wLogDate estimates in relation to: (1) the age of the node (Fig. 2a), (2) the length of the true branch in

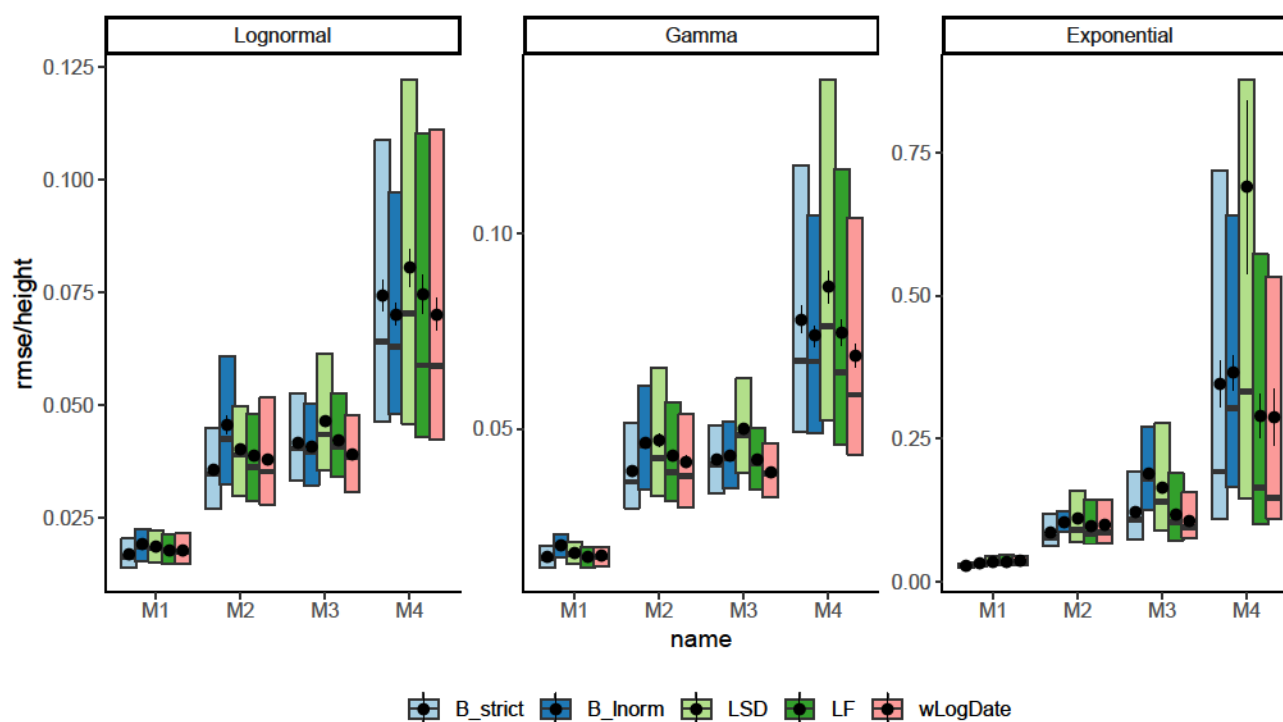


FIG. 3. Distributions of RMSE normalized by the tree height for internal node ages inferred by all methods on model trees M1–M4, each with clock models Lognorm, Gamma, and Exponential. Boxes show median, 10% and 90% quantiles; dots and error bars show mean and standard error (100 replicates).

time unit (Fig. 2b), and (3) the error of the branch lengths (in substitution unit) estimated by PhyML (Fig. S6). Overall, we do not observe a substantial change in the mean estimation error of wLogDate as the node age and the branch length change. The variance, however, can vary with node ages (Figure 2a), especially in M3 and M4 model conditions. Moreover, longer branches have a tendency to have higher variance in absolute terms (Fig. 2b). However, note that the relative error (i.e., log-odds error) dramatically *reduces* as branches become longer (Fig. S6). In studying the effect of the error in branch length estimation, we see that wLogDate can underestimate the branch time if the branch length in substitution unit is extremely underestimated (Fig.S6a, Supplementary Materials). In some

cases wLogDate under-estimates branch times by two order of magnitude or more; all of these cases correspond to super-short branches with substitution unit branch length underestimated by three or four orders of magnitude. As mentioned previously, extremely short estimated branch lengths are often the zero-event branches (Fig.S7), which are unavoidable for short sequences.

We next compare wLogDate to alternative methods, namely LF, LSD, and BEAST with strict-clock and Lognormal clock. Measured by RMSE, the accuracy of all methods varies substantially across model trees (M1 – M4) and models of rate variation (Fig. 3). Comparing methods, for many conditions, wLogDate has the lowest error, and in many others, it is ranked

second best (Table 1). Across all conditions, wLogDate has a mean rank of 1.75, followed by BEAST with strict clock with a mean rank 2; mean normalized RMSE of wLogDate, LF, BEAST-strict, BEAST-LogNormal, and LSD are 0.072, 0.074, 0.077, 0.087, and 0.116, respectively. Interestingly, in contrast to wLogDate, LSD seems to often underestimate branch times for many short branches even when they are estimated relatively accurately in substitution units (Fig. S6b, Supplementary Materials). For all methods, errors are an order of magnitude smaller for the LogNormal and Gamma models of rate variations compared to the Exponential model. In terms of trees, M4, which simulates inter-host evolution and high the largest height, presents the most challenging case for all methods. Interestingly, wLogDate has the best accuracy under all parameters of M4 tree and also all parameters of M3 (thus, both inter-host conditions). On M1, all methods have very low error and perform similarly (Fig. 3).

Among other methods, results are consistent with the literature. Despite its conceptual similarity to wLogDate, LSD has the worst accuracy. On M1 and M2, LSD is competitive with other methods; however, on M3 and M4, it has a much higher error, especially with the Exponential model of rate variation. With the LogNormal clock model, BEAST-LogNormal is better than BEAST-strict only for M4 but not for M1–M3; in fact, BEAST-LogNormal has the

highest error for the M2 condition. This result is surprising given the correct model specification. Nevertheless, BEAST-LogNormal is competitive only under the LogNormal model of rate variation and is one of the two worst methods elsewhere. Thus, BEAST-LogNormal is sensitive to model misspecification. In contrast, BEAST-strict is less sensitive to the model of rate variation and ranks among the top three in most cases. In particular, BEAST-strict is always the best method for intra-host ladder-like trees M1 and M2.

Accuracy of tMRCA follows similar patterns (Fig. 4). Again, the Exponential rate variation model is the most difficult case for all methods, resulting in biased results and highly variable error rates for most methods. In all conditions of M3 and M4, wLogDate has the best accuracy and improves on the second best method by 9 – 66% (Table 2). For M1 and M2, BEAST-strict is often the best method. The mean tMRCA error of wLogDate across all conditions is 4.83 (years), which is substantially better than the second best method, BEAST-strict (6.21).

In terms of the mutation rate, the distinction between methods is less pronounced (Table S1). wLogDate is the best method jointly with the two strict clock models BEAST-strict and LF. Overall, even though LF and wLogDate tend to over-estimate mutation rates, both have less biased results compared to other methods (Fig. 4). LSD and BEAST-LogNormal have the highest errors; depending on the condition, each can

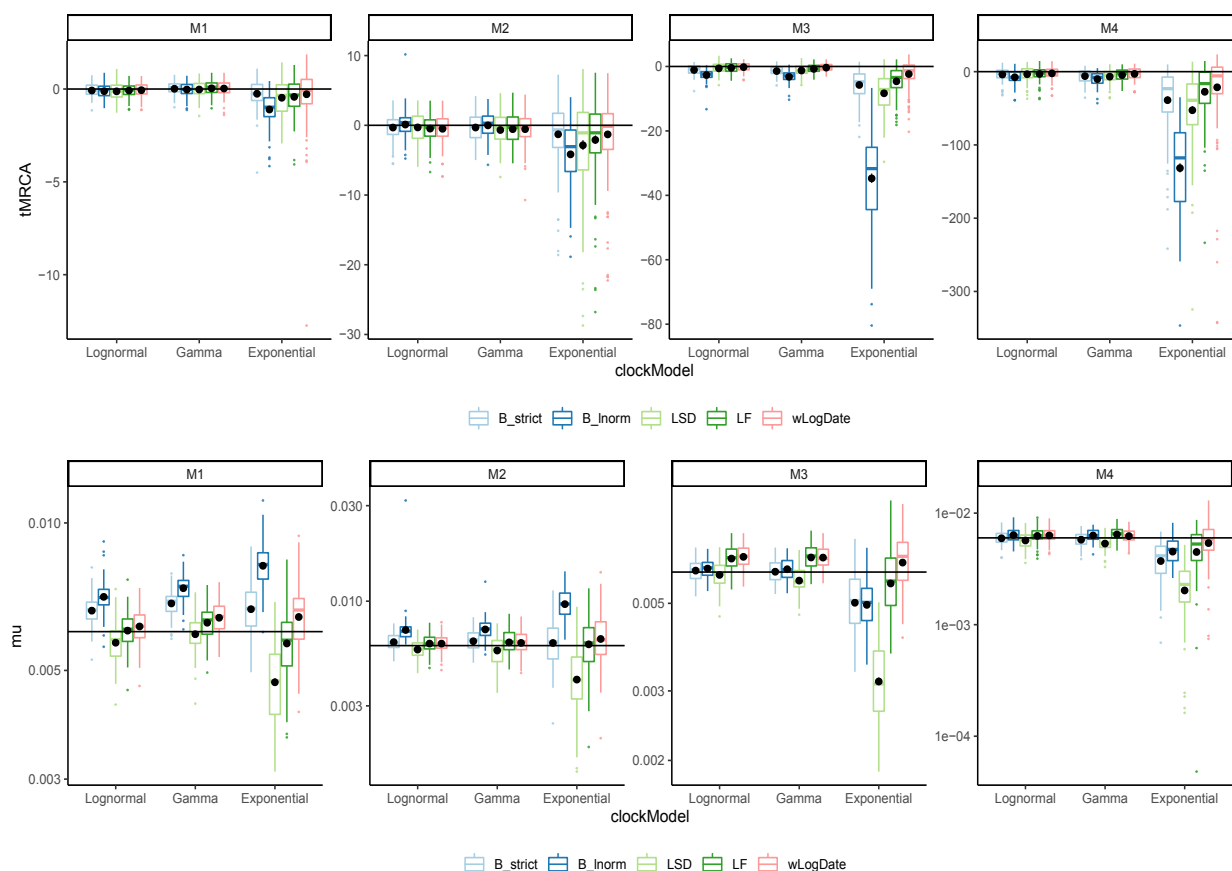


FIG. 4. The inferred (top) tMRCA and (bottom) expected mutation rate on different tree models and clock models. Distributions are over 100 replicates. The solid horizontal lines indicate the true mutation rate and tMRCA. Each black is the average of the inferred values for each method under each model condition. We remove 6 outlier data points (2 LF, 1 LSD, 2 BEAST-LogNormal, 1 BEAST-Strict) with exceptional incorrect tMRCA (< -350) in the M4/Exponential model.

Tree	Clock Model	B_strict	B_inorm	LF	LSD	RTT	wLogDate
M4	Lognormal	6.99	9.50	6.66	7.38	9.28	6.11 (9%↓)
	Gamma	7.83	10.48	7.02	8.48	8.24	6.28 (12%↓)
	Exponential	43.5	140.9	116.2	62.2	31.5	32.5 (3%↑)
M3	Lognormal	1.37	2.60	1.21	1.39	1.46	1.03 (17%↓)
	Gamma	1.60	3.14	1.23	1.67	1.42	0.97 (27%↓)
	Exponential	5.76	34.67	4.87	8.35	3.39	2.94 (66%↓)
M2	Lognormal	1.40	1.41	1.50	1.63	2.19	1.47 (5%↑)
	Gamma	1.54	1.44	1.75	1.92	2.56	1.66 (15%↑)
	Exponential	3.39	4.59	4.28	5.27	5.23	3.72 (10%↑)
M1	Lognormal	0.28	0.28	0.30	0.37	0.78	0.30 (7%↑)
	Gamma	0.27	0.29	0.32	0.35	0.80	0.30 (11%↑)
	Exponential	0.60	1.11	0.79	0.82	1.37	0.69 (15%↑)
Average		6.21	17.54	12.17	8.13	5.68	4.83

Table 2. Mean absolute error of the inferred tMRCA of BEAST_strict, BEAST_lognorm, LF, LSD, RTT, and wLogDate. For wLogDate, parenthetically, we compare it with the best (↑) or second best (↓) method for each condition. We show percent improvement by wLogDate, as measured by the increase in the error of the second best method (wLogDate or the alternative) divided by the error of the best method.

overestimate or underestimate the rate but LSD tends to underestimate while BEAST-LogNormal tends to overestimate. On M1, wLogDate and LF have a clear advantage over BEAST-strict, which tends to over-estimate the rate. On M2, the three methods have similar accuracy. For M3 and M4, BEAST-strict under-estimates the rate under the Exponential model of rate variation, and wLogDate and LF are closer to the true value. For all methods, M4 is the most challenging case.

We also compare confidence intervals obtained from wLogDate and BEAST (Fig.5). Although wLogDate intervals are on average 2.7 times larger

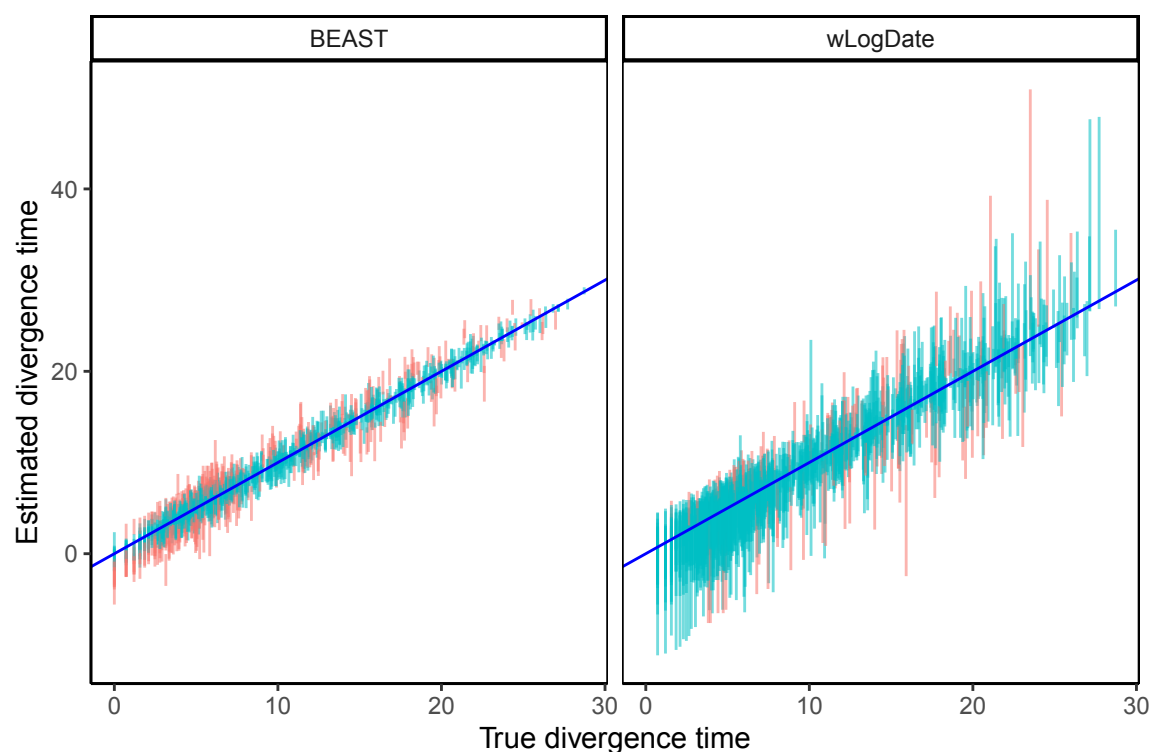


FIG. 5. Estimated versus true divergence time. Each bar corresponds to the 95% confidence interval (CI) of one node estimate (each of the 109 nodes of the 10 replicates) by BEAST strict clock and wLogDate. Red color is used to mark points where the true time falls outside the CI.

than BEAST, 33% and 12% of the true values fall outside the 95% confidence interval for BEAST and wLogDate, respectively. Thus, while both methods under-estimate the confidence interval range, wLogDate, with its larger intervals, is closer to capturing the true value in its confidence interval at the desired level.

Finally, we compared all methods in terms of their running time (Table S2). LSD and LF are the fastest methods in all conditions, always taking tens of seconds (less than a minute) on these data. The running time of wLogDate depends on the model condition and can be an order of magnitude higher for Exponential rates than the other two models of rate variation. Nevertheless, wLogDate finishes on average in half a minute to 12 minutes, depending on the model condition. In contrast,

BEAST took close to one hour with strict clock and close to two hours with the LogNormal model (and even more if run with longer chains; see Table S5 in Supplementary Materials).

Simulated data with autocorrelated rate

In simulations with the autocorrelated rate model, we compare wLogDate to LF and RelTime (Fig. 6 and Table S7) and wLogDate to DAMBE (Table S8). The distribution of the estimated divergence time of uncalibrated internal nodes does not show any sign of biased in divergence time estimation for either method. All methods seem to give less varied estimates for the younger nodes (i.e. those with higher divergence times) and have more varied estimates for older nodes. In addition, the estimates of wLogDate are more concentrated around the true values than that of LF and

RelTime, indicating a better accuracy. In two test cases (out of 200), LF had extremely high error (Fig.S7). Once those two cases are removed, the average RMSE normalized by tree height is 0.09 for wLogDate, 0.10 for LF, and 0.13 for RelTime (Table S7). Comparing to LF and wLogDate, RelTime gives wider distributions of the estimates for a large portion of the nodes. Finally, the comparison in running time of wLogDate and RelTime is shown in Fig.S8 (Supplementary Materials).

Comparing to DAMBE in inferring unit time trees, wLogDate has lower error in 6/10 replicates and DAMBE has lower error in the remaining 4 replicates (Table S8). Overall, the average error of wLogDate is 9.40%, which is slightly lower than that of DAMBE at 9.66%.

Biological data

On the H1N1 dataset, the best available evidence has suggested a tMRCA between December 2008 and January 2009 (Hedge *et al.*, 2013; Lemey *et al.*, 2009; Rambaut and Holmes, 2009). wLogDate inferred the tMRCA to be 14 December 2008 (Fig. 7a), which is consistent with the literature. LF and LSD both infer a slightly earlier tMRCA (10 November 2008), followed by BEAST-strict and BEAST-lognorm (October 2008 and July 2008), and finally BEAST runs using the phyML tree (Feb. 2008 for strict and July 2007 for LogNormal). While the exact tMRCA is not known on this real data, the results demonstrate that wLogDate, on a real

data, produces times that match the presumed ground truth.

On the HIV dataset, wLogDate inferred a tMRCA of 1958 with only a handful of lineages coalescing in the 1950s and most others coalescing in 1960s and early 1970s (Fig. S5). The recovered tMRCA is within the range postulated in the literature for subtype B (Gilbert *et al.*, 2007; Wertheim *et al.*, 2012) and the fact that randomly sampled HIV lineages across USA tend to coalesce deep in the tree is a known phenomenon. LF and LSD recovered the tMRCA of 1952 and 1953, respectively. Comparing to wLogDate, these two strict-clock methods postulate an earlier burst of subtype B (Fig. 7c). We were not able to run BEAST on this dataset.

On the Ebola dataset, the BEAST-1000 trees obtained from Dudas *et al.* (2017) inferred the tMRCA to be between 13 September 2013 and 26 January 2014 (95% credible interval) and the BEAST-MCC inferred the tMRCA to be 5 December 2013 as reported by Volz and Frost (2017). Here, wLogDate inferred a tMRCA on 7 December 2013, which is very close to the estimate by BEAST. Both LF and LSD inferred an earlier tMRCA: 29 October 2013 for LF and 2 October 2013 for LSD, but still within the 95 per cent credible interval of BEAST-1000. LTT plots showed a similar reconstruction by all methods for this dataset (Fig. 7d).

We also compare running times of dating methods on the three real biological datasets

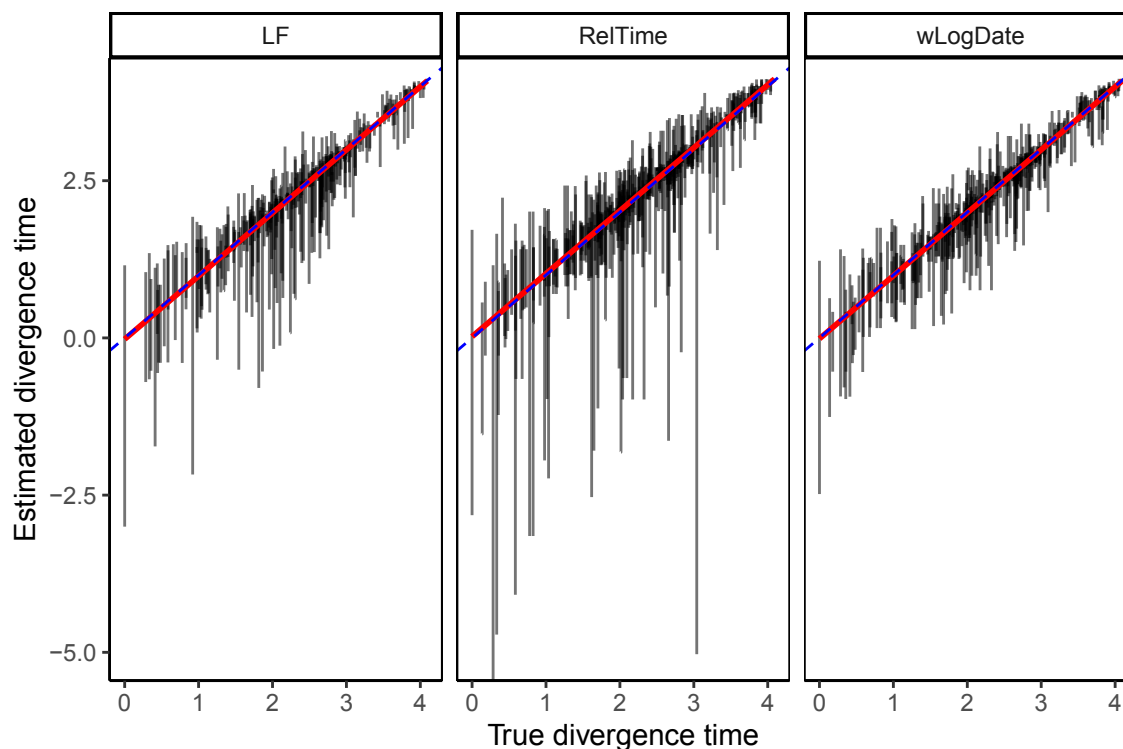


FIG. 6. Comparison of LF, RelTime, and wLogDate on the simulated data with autocorrelated rate model. The y-axis shows estimated divergence times of uncalibrated internal nodes while the x-axis shows the true divergence time. Each bar shows the 2.5% and 97.5% quantiles of the estimates of a single node's divergence time across 20 tests, each of them with different random choices of calibration points (thus, these are not CIs for one run). There are 10 replicate trees, each with 44 uncalibrated nodes (thus, 440 bars in total). This figure discards 2 tests (out of $10 \times 20 = 200$) where LF produced extremely erroneous time trees (see Fig. S9) for the full results). The root-mean-square error of the un-calibrated internal node ages, normalized by the tree height averaged across all replicates were 0.09, 0.1, and 0.13, respectively, for wLogDate, LF, and RelTime (see Table S7).

(Table S3). LSD was always the fastest, running in just seconds, compared to minutes for LF and wLogDate. LF is faster than wLogDate on the H1N1 and HIV data, while on Ebola data, wLogDate is faster. We report the running time for wLogDate as the sequential run of 10 independent starting points and note that wLogDate can easily be parallelized. We further tested the scaling of wLogDate with respect to the number of species by subsampling the HIV dataset to smaller numbers of species (Fig. S4). The results show that the running time of wLogDate increases slightly worse than quadratically with the increased number of species.

Discussion and future work

We introduced (w)LogDate, a new method for dating phylogenies based on a non-convex optimization problem. We showed that by log-transforming the rates before minimizing their variance, we obtain a method that performs much better than LSD, which is a similar method without the log transformation. In phylodynamics settings, our relatively simple method also outperformed other existing methods, including the Bayesian methods, which are much slower. The improvements were most pronounced in terms of the estimation of tMRCA and individual node ages and less so for the mutation rate. Moreover,

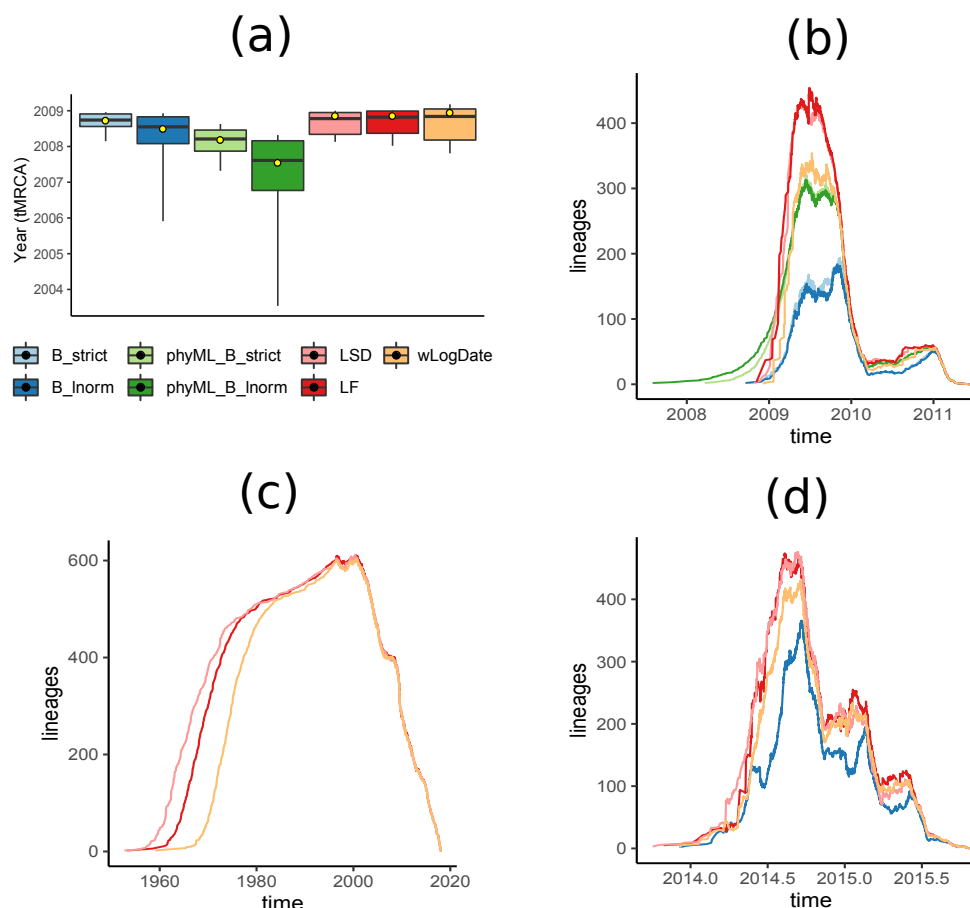


FIG. 7. (a) Inferred tMRCA of the H1N1 dataset. Boxplots represent the median, maximum, minimum, 97.5% and 2.5% quantiles of the bootstrap estimates for LF, LSD, and wLogDate, and of the posterior distribution for BEAST. Yellow dot shows the inferred tMRCA of the best ML or MAP tree. BEAST was run with 4 different settings: B_strict and B_inorm allow BEAST to infer both tree topology and branch lengths, with strict and LogNormal clock models; phyML_B_strict and phyML_B_inorm fixed the topology to the rooted phyML tree given to BEAST. All other methods (LSD, LF, and wLogDate) were run on the rooted phyML trees. Results for LSD, LF, and BEAST are all obtained from To *et al.* (2015). (b) LTT plot for all methods on the H1N1 data. (c) LTT plot of fast methods on the HIV dataset. (d) LTT plot of BEAST, LSD, LF, and wLogDate on the Ebola dataset.

improvements are most visible under the hardest model conditions, and are also observed in when data are generated according to autocorrelated model of rates.

The log transformation results in a non-convex optimization problem, which is harder to solve than the convex problems solved by LSD and LF. However, we note that the problem is convex for rate multipliers between 0 and e . In addition, given the advances in numerical methods for solving non-convex optimization problems, insistence on convex problems seems

unnecessary. Our results indicate that this non-convex problem can be solved efficiently in the varied settings we tested. The main benefits of the log transformation is that it allow us to define a scoring function that assigns symmetrical penalties for increased or decreased rates (Fig. 1a); as we argued, this symmetry is a desirable property of the penalty function for several clock models that deviate from a strict clock.

The accuracy of LogDate under varied conditions we tested is remarkable, especially given its lack of reliance on a particular model

of rate evolution. We emphasize that the parametric models used in practice are employed for mathematical convenience and not because of a strong biological reason to believe that they capture real variations in rates.

Even assuming biological realism of the rate model, the performance of the relaxed clock model used in BEAST was surprisingly low. For example, when rates are drawn from the LogNormal distribution, BEAST-strict often outperformed BEAST-LogNormal, especially in terms of the estimates of tMRCA and the mutation rate. We confirmed that the lower accuracy was not due to lack of convergence in the MCMC runs. We reran all experiments with longer chains (Table S5). to ensure ESS values are above 300 (Table S6). These much longer runs failed to improve the accuracy of the BEAST-LogNormal substantially and left the ranking of the methods unchanged (Fig. S10).

The LogDate approach can be further improved in several aspects. First, the current formulation of LogDate assumes a rooted phylogenetic tree, whereas most inferred trees are unrooted. Rooting phylogenies is a non-trivial problem and can also be done based on principles of minimizing rate variation (Mai *et al.*, 2017). Similar to LSD, LogDate can be generalized to unrooted trees by rooting the tree on each branch, solving the optimization problem for each root, and choosing the root that minimizes the (w)LogDate objective function. We leave the careful study of such an approach to the future work.

Beyond rooting, the future work can explore the possibility of building a specialized solver for LogDate to gain speedup. One approach could be exploiting the special structure of the search space defined by the tree, which is the strategy employed by LSD to solve the least-squares optimization in linear time. Divide-and-conquer may also prove effective.

The weighting scheme used in LogDate is chosen heuristically to deal with the deviations of estimated branch lengths from the true branch length. In future, the weighting schema should be studied more carefully, both in terms of theoretical properties and empirical performance.

We described, implemented, and tested LogDate in the condition where calibrations are given as exact times (for any combinations of leaves and internal nodes). While the current settings fit well to phylodynamics data, its application to paleontological data with fossil calibrations is somewhat limited due to the requirements for exact time calibrations (in contrast to the ability to allow minimum or maximum constraints on the ages, or a prior about the distribution of the ages as in BEAST and RelTime). While the mathematical formulation extends easily, treatment of fossil calibrations and uncertainty of times is a complex topic (Heath, 2012; Ho and Phillips, 2009) that requires the expansion of the current work. Applying LogDate for deep phylogenies would need further tweaks to the algorithm, including

changing equality to inequality constraints and the ability to setup feasible starting points for the solver.

In the studies of LogDate accuracy, we have explored various models for rate heterogeneity, including parametric models where rates are drawn i.i.d. from a fixed distribution (Log-normal, Exponential, and Gamma) and autocorrelated model where the rates of adjacent branches are correlated. Overall, none of the methods we studied is the best under all conditions. In phylodynamics data, our simulations showed that it is more challenging for all the dating methods to date the phylogenies of the inter-host evolution (M3 and M4) than the intra-host (M1 and M2). wLogDate outperforms other methods for the inter-host phylogenies, regardless of the model of rate heterogeneity. While all methods have lower error for intra-host trees, BEAST with strict-clock prior is often the best method. However, the differences between BEAST and wLogDate are small and wLogDate is often the second best. Thus, wLogDate works well for virus phylogenies, especially in inter-host conditions. Despite the fact that RelTime explicitly optimizes the rate for each pairs of sister lineages, wLogDate is more accurate than both LF and RelTime on the data where the rates are autocorrelated between adjacent branches. These results show that wLogDate is applicable to a fairly large number of models of the trees and the rates.

Nevertheless, the approach taken by wLogDate suffers from its own limitations. By including a single mean rate around which (wide) variations are allowed, wLogDate is expected to work the best when rates have distribution that are close to being unimodal. However, rates on real phylogenies may have sudden changes leading to bimodal (or multimodal) rate distributions with wide gaps in between modes. For example, certain clades in the tree may have local clocks that are very different from other clades. Such a condition has been studied by Beaulieu *et al.* (2015) for a dataset of seed plants. The authors setup a simulation where there are local clocks on the tree and the mean values of these clocks are different by a factor varying from 3 to 6. Beaulieu *et al.* (2015) point out that under such condition, especially when the rate shift occurs near the root, BEAST usually overestimates the time of the Angiosperm (i.e. gives older time) by a factor of 2 (BEAST results from Beaulieu *et al.* (2015) are reproduced in Fig. S11). We also tested wLogDate, LF, and RelTime on this dataset (Fig. S11). In scenario 2 of the simulation, where the rate shift between the two local clocks is extreme (a factor of 6), wLogDate clearly over-estimate the age of Angiosperms (by a median of 55%). In this same scenario, RelTime slightly underestimate the age (by 5%). In the other 4 scenarios where the rate shifts are more gentle, wLogDate continue to overestimate the age but by small margins (by 6%, 1%, 2%, and 5%), while RelTime underestimates

ages also by small margins (3%, 5%, 4%, 3%, and 3%). LF has similar patterns to wLogDate. These results point to a limitation of wLogDate (and the other dating methods) in phylogenies with multiple local clocks.

In addition to multiple clocks, future works should test LogDate under models where rate continuously change with time, and have a direction of change. Finally, to facilitate the comparison between different methods, we used the true topology with estimated branch lengths. Future work should also study the impact of the incorrect topology on LogDate and other dating methods.

Software availability. The LogDate software is available on <https://github.com/uym2/wLogDate> in open-source format. The command-line python tool is available through conda for easy installation. A link to a web sever making wLogDate available as a web-server is also available from the github page.

Data availability. All the data are available on <https://github.com/uym2/LogDate-paper>.

Acknowledgments

We thank Susan B. Little for providing the San Diego HIV sequence dataset used in this study.

This work was supported by the National Science Foundation (NSF) grant III-1845967 and a developmental grant from the University of California, San Diego Center for AIDS Research

(P30 AI036214), supported by the National Institutes of Health.

References

- Aban, I. B. and Meerschaert, M. M. 2004. Generalized least-squares estimators for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, 119(2): 341 – 352.
- Akerborg, O., Sennblad, B., and Lagergren, J. 2008. Birth-death prior on phylogeny and speed dating. *BMC Evolutionary Biology*, 8(1): 77.
- Beaulieu, J. M., O’Meara, B. C., Crane, P., and Donoghue, M. J. 2015. Heterogeneous Rates of Molecular Evolution and Diversification Could Explain the Triassic Age Estimate for Angiosperms. *Systematic Biology*, 64(5): 869–878.
- Bromham, L. and Penny, D. 2003. The modern molecular clock. *Nature Reviews Genetics*, 4(3): 216–224.
- Brown, R. P. and Yang, Z. 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evolutionary Biology*, 11(1): 271.
- Clifford, D., Cressie, N., England, J. R., Roxburgh, S. H., and Paul, K. I. 2013. Correction factors for unbiased, efficient estimation and prediction of biomass from log–log allometric models. *Forest Ecology and Management*, 310: 375 – 381.
- Donoghue, P. C. J. and Yang, Z. 2016. The evolution of methods for establishing evolutionary timescales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699): 20160020.
- Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7: 214.
- Drummond, A. J. and Suchard, M. A. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8(1): 114.
- Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5): e88.

- Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J. W., Duraffour, S., Elmore, M. J., Fakoli, L. S., Faye, O., Gilbert, M. L., Geva, S. M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D. S., Haagsmans, B. L., Hiscox, J. A., Jah, U., Kugelman, J. R., Liu, D., Lu, J., Malboeuf, C. M., Mate, S., Matthews, D. A., Matranga, C. B., Meredith, L. W., Qu, J., Quick, J., Pas, S. D., Phan, M. V. T., Pollakis, G., Reusken, C. B., Sanchez-Lockhart, M., Schaffner, S. F., Schieffelin, J. S., Sealfon, R. S., Simon-Loriere, E., Smits, S. L., Stoecker, K., Thorne, L., Tobin, E. A., Vand, M. A., Watson, S. J., West, K., Whitmer, S., Wiley, M. R., Winnicki, S. M., Wohl, S., Wölfel, R., Yozwiak, N. L., Andersen, K. G., Blyden, S. O., Bolay, F., Carroll, M. W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G. F., Garry, R. F., Goodfellow, I., Günther, S., Hap, C. T., Holmes, E. C., Kargbo, B., Keita, S., Kellam, P., Koopmans, M. P. G., Kuhn, J. H., Loman, N. J., Magassouba, N., Naidoo, D., Nichol, S. T., Nyenswah, T., Palacios, G., Pybus, O. G., Sabeti, P. C., Sall, A., Ströher, U., Wurie, I., Suchard, M. A., Lemey, P., and Rambaut, A. 2017. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650): 309–315.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.
- Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4): 406–416.
- Forest, F. 2009. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. *Annals of Botany*, 104(5): 789–794.
- Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7): 685–695.
- Gilbert, M. T. P., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M. 2007. The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Sciences*, 104(47): 18566–18570.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.
- Heath, T. A. 2012. A hierarchical Bayesian model for calibrating estimates of species divergence times. *Systematic biology*, 61(5): 793–809.
- Hedge, J., Lycett, S. J., and Rambaut, A. 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. *Biology Letters*, 9(5): 20130331.
- Hillis, D. M., Moritz, C., and Mable, B. K. 1996. *Molecular Systematics*, volume 2nd. Sinauer Associates.
- Ho, S. Y. 2014. The changing face of the molecular evolutionary clock. *Trends in Ecology & Evolution*, 29(9): 496 – 503.
- Ho, S. Y. W. and Phillips, M. J. 2009. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. *Systematic Biology*, 58(3): 367–380.
- Ho, S. Y. W., Duchêne, S., and Duchêne, D. 2015. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular Ecology Resources*, 15(4): 688–696.
- Huelsenbeck, J. P., Larget, B., and Swofford, D. 2000. A compound poisson process for relaxing the molecular clock. *Genetics*, 154(4): 1879–92.
- Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In *Mammalian protein metabolism, Vol. III (1969)*, pp. 21–132, volume III, pages 21–132.
- Keene, O. N. 1995. The log transformation is special. *Statistics in Medicine*, 14(8): 811–819.
- Kishino, H., Thorne, J. L., and Bruno, W. J. 2001. Performance of a Divergence Time Estimation Method

- under a Probabilistic Model of Rate Evolution. *Molecular Biology and Evolution*, 18(3): 352–361.
- Kodandaramaiah, U. 2011. Tectonic calibrations in molecular dating. *Current Zoology*, 57(1): 116–124.
- Kumar, S. 2005. Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6(8): 654–662.
- Kumar, S. and Hedges, S. B. 2016. Advances in Time Estimation Methods for Molecular Data. *Molecular Biology and Evolution*, 33(4): 863–869.
- Lalee, M., Nocedal, J., and Plantenga, T. 1998. On the implementation of an algorithm for large-scale equality constrained optimization. *SIAM Journal on Optimization*, 8(3): 682–706.
- Langley, C. H. and Fitch, W. M. 1974. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution*, 3(3): 161–177.
- Lemey, P., Suchard, M., and Rambaut, A. 2009. Reconstructing the initial global spread of a human influenza pandemic: A bayesian spatial-temporal model for the global spread of h1n1pdm. *PLoS currents*, 1: RRN1031–RRN1031. 20029613[pmid].
- Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. 2007. A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution*, 24(12): 2669–2680.
- Mai, U., Sayyari, E., and Mirarab, S. 2017. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLOS ONE*, 12(8): e0182238.
- Meaney, P. M., Fang, Q., Rubaek, T., Demidenko, E., and Paulsen, K. D. 2007. Log transformation benefits parameter estimation in microwave tomographic imaging. *Medical Physics*, 34(6Part1): 2014–2023.
- Nee, S., May, R. M., and Harvey, P. H. 1994. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309): 305–311.
- Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1).
- Pulquério, M. J. and Nichols, R. A. 2007. Dates from the molecular clock: how wrong can we be? *Trends in Ecology & Evolution*, 22(4): 180 – 184.
- Rambaut, A. and Grass, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3): 235–238.
- Rambaut, A. and Holmes, E. 2009. The early molecular epidemiology of the swine-origin a/h1n1 human influenza pandemic. *PLoS currents*, 1: RRN1003–RRN1003. 20025195[pmid].
- Rutschmann, F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity & Distributions*, 12(1): 35–48.
- Rzhetsky, A. and Nei, M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5): 1073–1095.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406–425.
- Sand, A., Holt, M. K., Johansen, J., Fagerberg, R., Brodal, G. S., Pedersen, C. N. S., and Mailund, T. 2013. Algorithms for computing the triplet and quartet distances for binary and general trees. *Biology*, 2(4): 1189–209.
- Sanderson, M. J. 1997. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Molecular Biology and Evolution*, 14(12): 1218–1231.
- Sanderson, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? In *Molecular systematics of plants II*, pages 242–264. Springer.
- Sanderson, M. J. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized

- Likelihood Approach. *Molecular Biology and Evolution*, 19(1): 101–109.
- Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2): 301–302.
- Schwartz, J. H. and Maresca, B. 2006. Do molecular clocks run at all? a critique of molecular systematics. *Biological Theory*, 1(4): 357–371.
- Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L., and Mullins, J. I. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12): 10489–502.
- Snir, S., Wolf, Y. I., and Koonin, E. V. 2012. Universal Pacemaker of Genome Evolution. *PLoS Computational Biology*, 8(11): e1002785.
- Stynes, D. J., Peterson, G. L., and Rosenthal, D. H. 1986. Log transformation bias in estimating travel cost models. *Land Economics*, 62(1): 94–103.
- Tamura, K., Battistuzzi, F. U., Billings-Ross, P., Murillo, O., Filipinski, A., and Kumar, S. 2012. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences*, 109(47): 19333–19338.
- Tamura, K., Tao, Q., and Kumar, S. 2018. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. *Molecular Biology and Evolution*, 35(7): 1770–1782.
- Tavaré, S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17: 57–86.
- Thorne, J. L., Kishino, H., and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12): 1647–1657.
- To, T.-H., Jung, M., Lycett, S., and Gascuel, O. 2015. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, 65(1): 82–97.
- Volz, E. M. and Frost, S. D. W. 2017. Scalable relaxed clock phylogenetic dating. *Virus Evolution*, 3(2): vex025.
- Volz, E. M., Koelle, K., and Bedford, T. 2013. Viral Phylodynamics. *PLoS Computational Biology*, 9(3).
- Wertheim, J. O., Sanderson, M. J., Worobey, M., and Bjork, A. 2010. Relaxed Molecular Clocks, the Bias–Variance Trade-off, and the Quality of Phylogenetic Inference. *Systematic Biology*, 59(1): 1–8.
- Wertheim, J. O., Fourment, M., and Kosakovsky Pond, S. L. 2012. Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy. *Molecular Biology and Evolution*, 29(2): 451–456.
- Xia, X. 2018. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Molecular Biology and Evolution*, 35(6): 1550–1552.
- Xia, X. and Yang, Q. 2011. A distance-based least-square method for dating speciation events. *Molecular Phylogenetics and Evolution*, 59(2): 342 – 353.
- Xiao, X., White, E. P., Hooten, M. B., and Durham, S. L. 2011. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology*, 92(10): 1887–1894.
- Zuckerandl, E. 1962. Molecular disease, evolution, and genetic heterogeneity. *Horizons in biochemistry*, pages 189–225.