

# The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates

Julien Soubrier,<sup>\*1</sup> Mike Steel,<sup>2</sup> Michael S.Y. Lee,<sup>3</sup> Clio Der Sarkissian,<sup>1</sup> Stéphane Guindon,<sup>4</sup> Simon Y.W. Ho,<sup>5</sup> and Alan Cooper<sup>\*1</sup>

<sup>1</sup>Australian Centre for Ancient DNA, School of Earth & Environmental Sciences, University of Adelaide, Adelaide, SA, Australia

<sup>2</sup>Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand

<sup>3</sup>South Australian Museum, North Terrace, Adelaide, SA, Australia

<sup>4</sup>Department of Statistics, University of Auckland, New Zealand

<sup>5</sup>School of Biological Sciences, University of Sydney, Sydney, NSW, Australia

**\*Corresponding author:** E-mail: julien.soubrier@adelaide.edu.au; alan.cooper@adelaide.edu.au.

**Associate editor:** Jeffrey Thorne

## Abstract

Molecular evolutionary rate estimates have been shown to depend on the time period over which they are estimated. Factors such as demographic processes, calibration errors, purifying selection, and the heterogeneity of substitution rates among sites (RHAS) are known to affect the accuracy with which rates of evolution are estimated. We use mathematical modeling and Bayesian analyses of simulated sequence alignments to explore how mutational hotspots can lead to time-dependent rate estimates. Mathematical modeling shows that underestimation of molecular rates over increasing time scales is inevitable when RHAS is ignored. Although a gamma distribution is commonly used to model RHAS, we show that when the actual RHAS deviates from a gamma-like distribution, rates can either be under- or overestimated in a time-dependent manner. Simulations performed under different scenarios of RHAS confirm the mathematical modeling and demonstrate the impacts of time-dependent rates on estimates of divergence times. Most notably, erroneous rate estimates can have narrow credibility intervals, leading to false confidence in biased estimates of rates, and node ages. Surprisingly, large errors in estimates of overall molecular rate do not necessarily generate large errors in divergence time estimates. Finally, we illustrate the correlation between time-dependent rate patterns and differential saturation between quickly and slowly evolving sites. Our results suggest that data partitioning or simple nonparametric mixture models of RHAS significantly improve the accuracy with which node ages and substitution rates can be estimated.

**Key words:** among-site rate variation, molecular clock, substitution rate, divergence times, time-dependent rates.

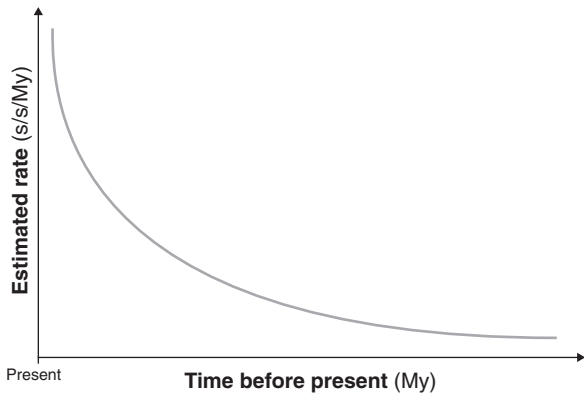
## Introduction

Increasingly large genetic datasets are becoming available to estimate the pattern and timing of evolutionary divergences among organisms. However, since the molecular clock hypothesis was first proposed in the 1960s (Zuckerkandl and Pauling 1962), numerous concerns have been raised about the estimation of molecular rates (Kumar 2005; Pulquerio and Nichols 2007). Considerable methodological improvements have also been made, such as the implementation of more biologically realistic yet computationally tractable models that accommodate changes in substitution rates along lineages, also known as “relaxed clocks” (e.g., Thorne et al. 1998; Huelsenbeck et al. 2000; Drummond et al. 2006).

In recent years, it has become apparent that substitution rate estimates are strongly related to the timescale over which they are measured and appear to vary systematically with the time point(s) chosen to calibrate a phylogenetic analysis (Parsons et al. 1997; Lambert et al. 2002; Howell et al. 2003; Ho et al. 2005, 2011). For a given locus or taxon, rate estimates for short time periods (e.g., within species) are typically far

higher than rate estimates for long time periods (e.g., between species and higher taxa) (Ho and Larson 2006). This time-dependent pattern (shown in [fig. 1](#)) has important consequences for the estimation of evolutionary timescales using molecular data. For example, the use of the human–chimpanzee divergence date to calibrate estimates of the time-scale of human dispersals has been shown to produce underestimates of the molecular rate, leading to overestimates of the age of migration events (Ho and Endicott 2008; Henn et al. 2009; Soares et al. 2009). Accordingly, it is recommended that calibration points should be selected according to the time-scale of interest (Ho et al. 2008) or that a correction be applied to estimates of mutation rates (Soares et al. 2009; Gignoux et al. 2011).

Several potential explanations have been suggested for time-dependent rates. For example, purifying selection (Ho et al. 2005; Woodhams 2006; Subramanian et al. 2009) and demographic events (Henn et al. 2009; Balloux and Lehmann 2011) have both been shown to contribute to such a pattern, but without individually being able to explain the full extent of time-dependent rates. A recent review summarized a wide



**Fig. 1.** Plot showing the time-dependent pattern of molecular rates. Rate estimates (in substitution per site per million year, s/s/My) decrease proportionally to the time period over which they are estimated.

range of possible contributing factors to this phenomenon including natural selection, calibration errors, model misspecification, and a number of other artifacts (Ho et al. 2011). The authors concluded that untangling all these factors concurrently would be a complex exercise, and thus making it difficult to correct explicitly for time-dependent rate estimates. Therefore, specific studies are needed to characterize and quantify the impact of each potential cause.

In this study, we analyse the effect of mutational hotspots, and more generally the impact of strong rate heterogeneity among sites (RHAS), on estimates of overall substitution rates and divergence times. It is well known that some positions of an alignment evolve at a substantially faster rate than others, and the impact of this among-site rate variation on phylogenetic reconstruction have been clearly demonstrated (Yang 1996; Simon et al. 2006). For example, hotspot positions have been described in the mitochondrial genome, which is one of the most commonly used genetic markers, for a wide range of organisms. Mitochondrial site-specific rates are well described and exhibit up to 1,000-fold difference between the fastest and slowest evolving positions (Galtier et al. 2006; Kjer and Honeycutt 2007; Rosset et al. 2008; Song et al. 2010). Interestingly, these mutational hotspots are distributed throughout the mitochondrial genome (including protein-coding genes, RNA genes, and noncoding regions) and are not limited to the least constrained parts such as the D-loop (Meyer and von Haeseler 2003). However, while RHAS has been put forward as a potential explanation for the time dependence of rates (Ho et al. 2005, 2007; Galtier et al. 2006; Henn et al. 2009), this has not been validated theoretically and the effects on molecular estimates of divergence times remain poorly explored.

The most common approach to account for RHAS while performing phylogenetic inference is to model site-specific rates using a gamma distribution. In such cases, the continuous distribution of site-specific rates is discretized into a predefined number of rate categories where the mean is set to 1.0 and the variance characterizes the rate variability. The probability of each site being associated with a rate category and the shape parameter of the distribution are then

estimated from the data (Yang 1994). However, more general mixture-models have recently been shown to outperform the discrete gamma model, particularly when multi-modal heterogeneity is present (Lartillot and Philippe 2004; Pagel and Meade 2004; Huelsenbeck and Suchard 2007). These mixture models do not involve the use of a parametric framework, which means that the estimated site-specific relative rates and frequencies are not constrained by any prespecified distribution. For example, Huelsenbeck and Suchard (2007) proposed the use of a Dirichlet process prior, which, unlike the discrete gamma model, does not assume a fixed number and distribution of rate classes.

Using theoretical arguments and Bayesian analyses of simulated sequences, we demonstrate that among-site rate heterogeneity by itself can produce a time-dependent pattern in overall rate estimates. Furthermore, we describe the consequences of such bias on age estimates obtained using phylogenetic methods. We also demonstrate that, contrary to expectations, increased mutational saturation does not invariably lead to increasingly biased estimates of divergence times.

## Materials and Methods

### Mathematical Model

Mathematical analyses were performed to analyse the effects of different models of among-site rate variation (e.g., a gamma distribution or a single-rate model) on the estimation of overall substitution rates, when sequences evolve under a multimodal rate model that allows for “hotspots” and other rate classes. The “equal input” substitution model has been chosen for convenience. For four-state sequences (e.g., DNA sequences), this model is termed the Felsenstein 1981 (or Tajima-Nei) model (Felsenstein 1981). It is more general than the Jukes-Cantor model (Jukes and Cantor 1969) because it allows base frequencies to be arbitrary (thus it has three parameters for nucleotide data or 19 for amino acid data), yet the model is simple enough to allow an exact mathematical analysis.

Let  $\kappa_t$  denote the evolutionary distance (the expected number of substitutions, per site) between the two sequences that are separated by time  $t$ . If sites undergo substitution at a rate that is constant with time (but possibly variable across sites) then  $\kappa_t = \mu t$ , where  $\mu$  is the average rate per site. More generally, if the instantaneous substitution rate depends on time, and if  $r(s)$  denotes the instantaneous substitution rate (averaged across sites) at time  $s$ , then  $\kappa_t = \int_0^t r(s) ds$ . Thus, by the fundamental theorem of calculus, the evolutionary rate (averaged across sites) at any time is given by

$$r(t) = \frac{d\kappa_t}{dt}. \quad (1)$$

Consider the probability  $p$  that a site has different states in the two sequences (i.e., the expected proportion of sites in different states). If there is no variation of rates across sites (though the rate may vary with time) the relationship between  $p$  and  $\kappa_t$  is given by

$$p = \gamma[1 - \exp(-\kappa_t/\gamma)],$$

where  $\gamma = 1 - \sum_i \pi_i^2$ , for the frequencies  $\pi$  of the states (see e.g., Semple and Steel 2003). For instance, for the Jukes-Cantor model, for which  $\pi_i = \frac{1}{4}$  for all four bases, we have  $\gamma = 3/4$ , and  $p = \frac{3}{4}[1 - \exp(-\frac{4}{3}\kappa_t)]$ .

Now, suppose that sites evolve at various rates (e.g., at one extreme, each site might have its own intrinsic rate, or the sites might fall into classes of given rates determined by sequence position, structure, or function). We refer to this as the underlying “generating process” denoted by  $G$ . We assume that the overall substitution rate of this generating process does not vary with time (it only varies across sites) and so for this process, ( $G$ ) the function  $r(t)$  given by equation (1) is constant. Let  $\mu$  denote the average substitution rate across sites,  $0 \leq r_1 \leq \dots \leq r_n$  denote relative substitution rates (i.e., absolute rate divided by  $\mu$ ), and  $\alpha_i > 0$  denote the proportion of sites that evolves at rate  $i$ . Thus,  $\alpha_i$  proportion of sites evolves at absolute rate  $\mu r_i$ , and  $\sum_{i=1}^n \alpha_i = 1 = \sum_{i=1}^n \alpha_i r_i$ .

Under this mixture generating process, the expected proportion  $p$  of sites in different states between the two sequences is given by  $p = g(t)$  where

$$g(t) := \sum_{i=1}^n \alpha_i \gamma [1 - \exp(-\mu r_i t / \gamma)]. \quad (2)$$

Notice that, for the generating process, the evolutionary distance is just  $\kappa_t = \mu t$ , and so we can rewrite equation (2) as

$$p = \sum_{i=1}^n \alpha_i \gamma [1 - \exp(-\kappa_t r_i / \gamma)]. \quad (3)$$

Now, suppose sequences evolve under the generating process ( $G$ ), but we use the sequence dissimilarities to estimate evolutionary distance by correcting them according to a second model  $M$ . We will refer to  $M$  as the “correction model.” We will assume that  $M$  is the same as  $G$  except for the way rates across sites are distributed.

If we estimate the evolutionary distance  $\kappa_t$  between two sequences by taking the proportion  $\hat{p}$  of sites at which those two sequences differ and then correct  $\hat{p}$  according to some  $M$ , then we will let  $\hat{\kappa}_t$  denote this estimated evolutionary distance. We will mostly assume that the correction model is a mixture of discrete rates, because implementations of continuous distributions such as the Gamma distribution typically treat this as a discrete (binned) distribution; however, the results that follow have a parallel treatment in the continuous case in which weighted sums are replaced by integrals with respect to the distribution of rates across sites, as described further in the Appendix.

In a discrete correction model, if  $r'_1, \dots, r'_{n'}$  denote the relative rates and  $\alpha'_i > 0$  their frequencies, then we can estimate  $\hat{\kappa}_t$  from  $\hat{p}$  by solving the equation:

$$\hat{p} = f(\hat{\kappa}_t) := \sum_{i=1}^{n'} \alpha'_i \gamma [1 - \exp(-\hat{\kappa}_t r'_i / \gamma)]. \quad (4)$$

Note that the number of classes in the correction model  $M$  can differ from the generating process  $G$ . Also, the  $r'_i$  values

are relative rates and so we have  $\sum_{i=1}^{n'} \alpha'_i r'_i = 1$ , as before for  $G$ . Allowing  $r_1 = 0$  or  $r'_1 = 0$  (or both) allows for a class of invariable sites in the generating or correcting model, respectively.

Now, suppose sequence data are generated by the original generating process  $G$ . As the sequence length increases, the observed proportion of sites at which two sequences differ  $\hat{p}$  will converge to its expected value  $p$  determined by equation (3). Thus, from equation (4), the estimated evolutionary distance  $\hat{\kappa}_t$  between any two sequences obtained by using the correction model  $M$  will converge (as the sequence length grows) to the solution of the equation:

$$f(\hat{\kappa}_t) = g(t). \quad (5)$$

We refer to the graph of  $\hat{r}(t) = \frac{d\hat{\kappa}_t}{dt}$  against  $t$  as the (estimated) “rate curve.” By equation (1),  $\hat{r}(t)$  describes how the instantaneous substitution rate must vary in the correction model  $M$  in order for the expected sequence dissimilarities to match their expected values under the process  $G$  that generated the sequences (and for which the substitution rate does not change with time). Note that  $\hat{r}(t)$  differs from the associated rate function for  $G$ , which is simply  $r(t) = \mu$  (eq. 1), because  $G$  and  $M$  have different distribution of rates across sites. This difference will usually lead to a nonlinear dependence of  $\hat{\kappa}_t$  on  $t$  via equation (5), and thus a time-varying derivative (i.e.,  $\hat{r}(t)$  will vary with time, unlike  $r(t)$ ).

In a discrete correction model, we have from equation (4):  $f(x) = \sum_{i=1}^{n'} \alpha'_i \gamma [1 - \exp(-x r'_i / \gamma)]$ , which defines  $\hat{\kappa}_t$  as a function of  $t$  as the solution to the equation  $\hat{\kappa}_t = f^{-1}[g(t)]$  (provided this solution exists). Thus, the associated rate curve is given by differentiating both sides of equation (5) to obtain  $f'(\hat{\kappa}_t) \cdot \frac{d\hat{\kappa}_t}{dt} = g'(t)$  (where  $f'(*)$  denotes the derivative of  $f$  with respect to  $*$ ). Consequently, from equation (1), we have

$$\hat{r}(t) = \frac{g'(t)}{f'(\hat{\kappa}_t)} = \frac{g'(t)}{f'[f^{-1}[g(t)]]}. \quad (6)$$

This flexible mathematical framework allows exploration of overall rates over time for a wide range of generating processes ( $G$ ) and correction models ( $M$ ). For example, suppose the correction model ( $M$ ) assumes a single rate process. In this case, we have  $f(x) = \gamma[1 - \exp(-x/\gamma)]$ ,  $f'(x) = \exp(-x/\gamma)$  and  $f^{-1}(x) = -\gamma \log(1 - x/\gamma)$ . Thus,  $f'[f^{-1}(x)] = 1 - x/\gamma$  and so equation (6) becomes

$$\hat{r}(t) = \frac{g'(t)}{1 - g(t)/\gamma}. \quad (7)$$

To specialize further, suppose that sites are generated under a two-rate process (90% of sites at a “slow” rate 0.1 and 10% at a “fast” rate 1.0) but that the correction model assumes a single rate process. The average substitution rate in the generating process is  $\mu = 0.9 \times 0.1 + 0.1 \times 1.0 = 0.19$ , and equation (7) gives

$$\hat{r}(t) = \mu \cdot \frac{0.9r_1 \exp(-\mu r_1 t / \gamma) + 0.1r_2 \exp(-\mu r_2 t / \gamma)}{0.9 \exp(-\mu r_1 t / \gamma) + 0.1 \exp(-\mu r_2 t / \gamma)}.$$

In this formula,  $r_1$  and  $r_2$  are the relative rates ( $r_1 = 0.1/0.19$ ,  $r_2 = 1.0/0.19$ ) and the rate curve starts at  $\mu = 0.19$ , but decays toward 0.1 as  $t$  grows. This rate curve is shown in figure 2 (curve III); notice that the parameter  $\gamma$  only affects the scale of the time axis and not the shape, intercept or asymptote of the rate curve.

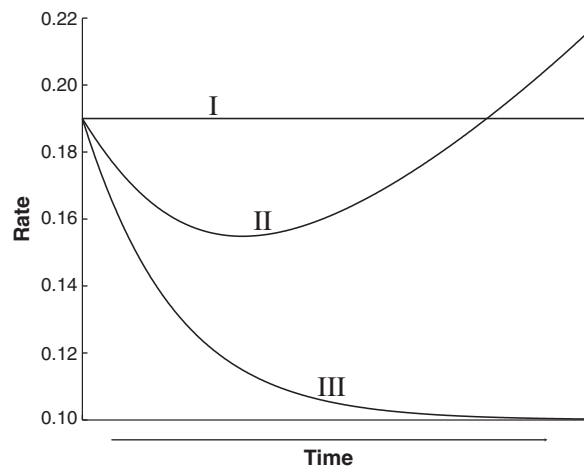
### Simulations

Simulated DNA datasets were used to explore the impact of RHAS on inferred rates and node ages. Given the results from the mathematical exploration and the potentially large parameter space, the proportion of fast sites was fixed at 10% for the two main scenarios (table 1). The first scenario, Scenario 1, has fast substitution rates overall (rates of 100 and 0.1 substitutions/site/My) and a large ratio between fast and slow rates (1,000:1). This extreme case of RHAS was chosen to exacerbate the phenomenon of interest, making the consequences more readily quantifiable. Scenario 2 has slower substitution rates overall and a smaller ratio between fast and slow rates (1.0 and 0.01 substitutions/site/My; rate ratio of 100:1). This second scenario represents more biologically realistic rate heterogeneity, being similar to that observed in, for example, mitochondrial sequences (supplementary material, Supplementary Material online).

Nucleotide sequence evolution was simulated on known trees using a strict molecular clock. First, 10-taxon trees were simulated by sampling from a coalescent prior, with constant population size of 100,000 individuals, using the program BEAST v1.6 (Drummond and Rambaut 2007). The ages of four nodes of interest were constrained to 5, 10, 20, and 50% of the total tree height (as shown in fig. 6B). The remaining portion of each tree was generated randomly under the coalescent.

Each tree was then rescaled to 15 different heights, with root ages ranging from 1,000 years to 50 My (1, 2, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000, 5,000, 10,000, 20,000, and 50,000 thousand years) for Scenario 1, and from 5,000 years to 200 My for Scenario 2. Scaling the same phylogenetic tree for each of the 15 analyses enables us to evaluate directly the influence of timescale on estimates of overall rates and node ages. Setting the root of each tree at different depths allows us to study the potential time dependence of these parameters.

Simulations from Scenario 1 and Scenario 2 were replicated 10 times (i.e., using 10 different trees). The constant-size coalescent prior, from which the starting trees were sampled, models the evolutionary process at the population level. This model is biologically relevant for relatively short time-scales, but is less realistic over some of the longer time periods studied here. Therefore, we based an additional replicate on trees generated using the Yule process, which offers a more realistic model of the branching process at the species level (supplementary material, Supplementary Material online). Considering the simplicity of the simulation scenarios (with only fast- and slow-evolving sites with a large difference between the two rates), we performed additional simulations that included sites with an intermediate substitution rate and a reduced rate ratio. In this particular scenario, 25% are



**Fig. 2.** Mathematical exploration of the impact of various correction models on inferred rates when the generating model considers a proportion of fast sites of 10% and a fast:slow rate ratio of 1:0.1. Correction using the true generating model leads to a constant rate curve (curve I). However, correcting with a model that assumes only a single rate class leads to a rate curve that declines toward the lowest of the two rates, that is, 0.1 (curve III). Intermediate between these is the rate curve when correcting using a continuous gamma distribution with shape parameter 1, leading to the “U-shaped” (curve II). Further decrease of the gamma distribution shape parameter results in a curve that increases for all  $t > 0$ . See text (and Appendix) for further details.

**Table 1.** Details about the Two Scenarios Used for Sequence Simulation.

	Rate (s/s/My)		Rate Ratio
	Fast Sites	Slow Sites	
Scenario 1	100	0.1	1000:1
Scenario 2	1	0.001	100:1

“fast” sites evolving at 1.0 substitutions/site/My, 25% are “intermediate” sites at 0.5 substitutions/site/My, and 50% are slow sites at 0.1 substitutions/site/My (giving a fast:intermediate:slow rate ratio of 10:5:1).

We used Seq-Gen (Rambaut and Grassly 1997) to simulate sequence evolution under the HKY85 model of nucleotide substitution (Hasegawa et al. 1985) with a transition/transversion ratio of 1 ( $\kappa = 2$ ) and equal nucleotide frequencies. Two subsets of data were simulated for each alignment, one comprising 9,000 slow-evolving sites and another comprising 1,000 fast-evolving sites, using the substitution rates described earlier for each scenario. The two subsets were concatenated to produce alignments of 10,000 sites.

Sequence alignments were analysed using BEAST v1.6. Based on the simulated values, we chose to place a strong narrow prior on the root height to remove the effects of calibration uncertainty. This was done by specifying a normal prior centered on the simulated root height, with a small standard deviation (0.1% of the mean). The substitution model and the ratio of transitions to transversions were chosen to match those used in the simulation process. Each

analysis was performed with and without a model allowing gamma-distributed rates among sites, with four discrete rate categories. The length of the Markov chain Monte Carlo was set to 100,000,000 steps, with samples drawn every 10,000 steps. The results were processed with Tracer v1.5 (Rambaut and Drummond 2007) to check that each sampled parameter had an effective sample size over 100. The default 10% burn-in was adjusted by hand where necessary to ensure sampling from the stationary distribution only.

To replicate the analysis performed with BEAST, we used the software PhyTime (from the PhyML package, Guindon 2010), which implements a Bayesian approach to estimate divergence times on a fixed tree topology. PhyTime was applied for Scenario 1, both with and without gamma-distributed rates among sites ([supplementary material, Supplementary Material](#) online).

In addition, PhyTime allowed the implementation of a mixture model in which the relative rates and the corresponding class frequencies are directly estimated from the data rather than being forced to conform to a parametric distribution such as the discrete gamma. We term this model the “FreeRate model,” and two classes of rate were defined a priori to fit the simulation scenario. More precisely, when taking a parametric approach by using a discrete gamma distribution for modeling RHAS, values of the relative rates and the corresponding frequencies are all determined by the shape parameter of the gamma density. Hence, characterizing the variability of rates among sites amounts to estimating this shape parameter. A simple extension to this model involves abandoning the parametric framework. We assume instead that the relative rates and frequencies of the mixture model do not have to match any pre-specified distribution. Each rate and frequency is then directly estimated from the data. While such approach comes at the price of an increase in the number of parameters to estimate compared with the gamma distribution (one for gamma against  $2c-2$  for the non parametric technique with  $c$  classes of relative rates), it provides a more flexible description of the RHAS.

We give below a more detailed description of the FreeRate model of RHAS. The likelihood at site  $s$  given the tree topology  $\tau$ , a vector of branch lengths  $L$  and a rate matrix  $Q$  is

$$p(D_s|\tau, L, Q) = \sum_{i=1}^c p(D_s|\tau, L \times r_i, Q) f_i,$$

where  $r_i$  is the relative rate for class  $i$  and  $f_i$  is the corresponding frequency. We have  $\sum f_i = 1$  and  $\sum f_i r_i = 1$ . Bayesian estimation of  $f_i$ s and  $r_i$ s requires to specify the joint prior  $p(f, r)$ , where  $f$  and  $r$  are the vectors of  $f_i$  and  $r_i$  values, respectively. In its current implementation in PhyTime, this prior is uniform.

In practice, sampling values of  $f$  and  $r$  from their posterior density cannot be done by applying a standard Metropolis-Hastings step to each element in these vectors successively. Indeed, changing one value in  $f$  or  $r$  would violate the two constraints  $\sum f_i = 1$  and  $\sum f_i r_i = 1$ . A solution here is to introduce new variables. We first have

$f_i = (y_{(i)} - y_{(i-1)})/y_{(c)}$  where  $y_{(i)}$  is the  $i$ th order statistic of  $y$  (with  $i = 1 \dots c$  and  $y_{(0)} = 0$ ) and the  $y_i$ s are independent and identically distributed random variables. The second variable,  $z$ , is such that  $\sum_{i=1}^c f_i z_i = K$ , where the  $z_i$ s are independent and identically distributed random variables and  $r_i = z_i/K$ . Each element in  $y$  and  $z$  is then updated successively using Metropolis-Hastings. The Hastings ratio for the proposal on the variables of interests, that is,  $f$  and  $r$  is derived by accounting for the change of variables between  $y$  and  $f$  and between  $z$  and  $r$ .

### Saturation Analysis

Mutational saturation in the simulated sequences was assessed for the different root ages by plotting observed (uncorrected pairwise) against inferred genetic distances (Philippe et al. 2004). The uncorrected distances were obtained using the Dset command in PAUP 4.0b10 (Swofford 2003). Inferred distances were obtained using the APE module in R (Paradis et al. 2004; R Development Core Team 2011) to extract patristic distances from phylogenetic trees. The trees were inferred with the software PhyML (Guindon and Gascuel 2003), using the HKY85 substitution model. To quantify the degree of saturation, a linear regression model with zero intercept was fitted for each plot, the slope representing the degree of saturation (Jeffroy et al. 2006). The slopes for the concatenated subsets, and both the fast and slow site subsets, were compared through time (all plots are given in the [supplementary material, Supplementary Material](#) online).

## Results and Discussion

### Mathematical Model

The mathematical model was used to show the time-dependent rate pattern when the correction model differed from the generating model, and parameter exploration shows the limitations of the correction models based on a gamma distribution ([fig. 2](#)). More specifically, Curve I in [figure 2](#) illustrates the case where the generating process  $G$  is the same as the correction model  $M$  [note that in equation (7)  $\hat{r}(0) = \mu$ ], then  $g(t) = f(\mu t)$  and  $\hat{c} = \kappa = \mu t$  so  $\hat{r}(t) = \mu$ , as expected [this can be seen directly from equation (1) or derived from equation (6)]. In contrast, if a single rate is assumed for the correction model but the sequences are actually generated under a F81 model with 90% of the sites slow (rate 0.1) and 10% of the sites fast (1.0), then  $\hat{r}(t)$  starts at the rate 0.19 ( $= \mu$ ) and drops toward 0.1 as time increases (Curve III in [fig. 2](#); as described above).

These examples are part of a more general result, given as Theorem 1 below, that describes the behavior of the rate curve for small and larger values of  $t$ . The results demonstrate that the rate curve can take on a variety of shapes, depending on both the variation in rate classes and the smallest rate classes in  $G$  and  $M$ . In particular, not only can the rate curve change toward a new asymptote, but also it can be a “U-shaped” curve ([fig. 2](#), curve II) or a curve that initially decreases then increases toward a higher asymptote (or, conversely, initially increases and then approaches a lower

asymptote). These results are summarized in the theorem below, with the formal proof provided in the Appendix.

### Theorem 1

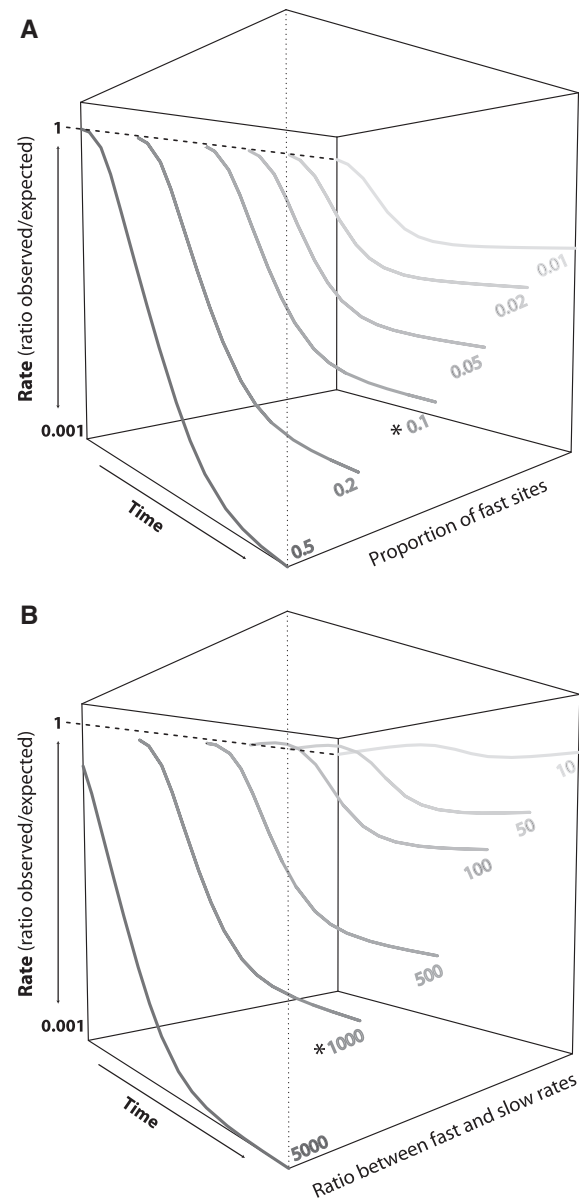
- a) The rate curve  $\hat{r}(t)$  initially decreases (i.e., its slope is negative at  $t = 0$ ) if and only if the variance of the relative rates of the generating process is greater than the variance of the relative rates of the correcting model.
- b) As  $t$  continues to increase, the rate curve  $\hat{r}(t)$  converges to some nonzero rate, decays to zero, or tends to infinity, depending on the relative rates of the generating process  $G$  and the correction model  $M$ . In particular, as  $t$  grows:
  - i) If  $G$  has more than one rate class and  $M$  has a single rate class, then the rate curve converges to the smallest nonzero absolute rate in  $G$ ;
  - ii) More generally, if both  $G$  and  $M$  have discrete rate classes and no invariable sites, then the rate curve converges to  $\mu r_1 / r'_1$ ;
  - iii) If  $G$  has discrete rate classes, and no invariable sites, and if the rates in  $M$  follow a continuous gamma distribution, then the rate curve tends to infinity;
  - iv) If  $G$  or  $M$  (or both) has invariable sites, then if there are fewer invariable sites in  $M$  than in  $G$ , the rate curve tends to zero; while if there are more invariable sites in  $M$  than  $G$ , then the rate curve tends to infinity as  $t$  approaches a sufficiently large (but finite) value.

Part (a) shows immediately why the two-rate mixture generating function will always lead to an initial decline in the rate curve when corrected under a single class (as in the earlier example), since the relative rates have positive variance in the generating process but zero variance in the correction model. Note also that a comparison of claims (ii) and (iii) of Part (b) reveals that a continuous gamma correction model behaves fundamentally differently from a discretized version (regardless of the number of bins) at sufficiently large time scales. Part (iv) of (b) also shows that infinite rates can be achieved at finite values of  $t$ , and beyond this vertical asymptote, equation (5) has no solution because there are simply insufficient variable sites available in the correction model to match the generating process.

We then used the mathematical model to explore the impact of changes in the proportion of fast sites, and in the ratio of fast to slow rates, on the time dependence of overall rates, when no correction is used to account for RHAS (fig. 3). The time-dependent underestimate of the rate is shown to be more pronounced both when the proportion of hotspots increases (fig. 3A), and when the ratio between fast and slow rates increases in the data (fig. 3B).

### Simulations

To understand the impact of among-site rate heterogeneity on the time dependence of molecular rates, we first discuss

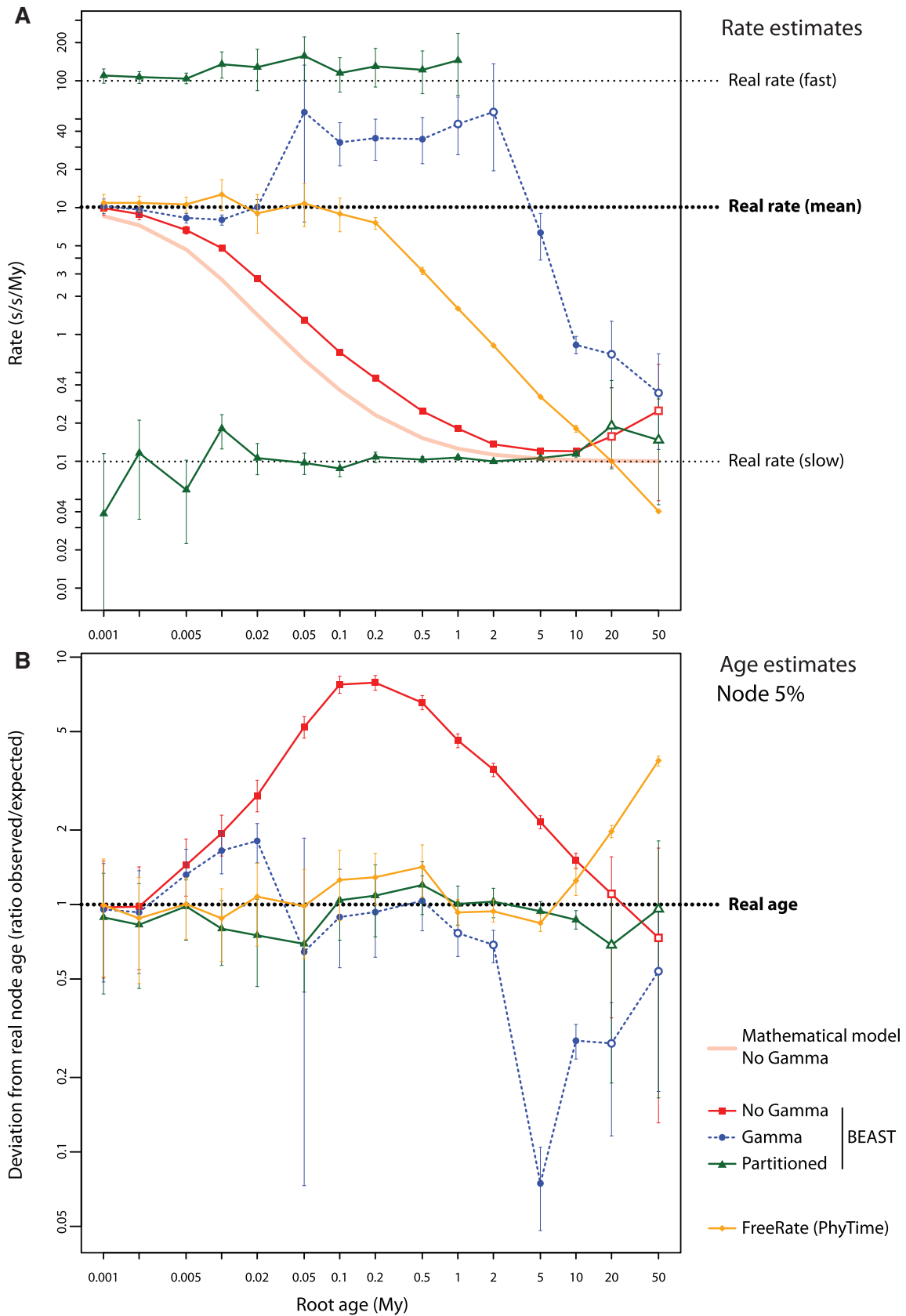


**FIG. 3.** Mathematical exploration of the relationship between divergence time, actual substitution rate, and inferred rate from the mathematical model. (A) Rates for various proportions of fast sites, when the ratio of fast to slow rates is 1,000. (B) Rates for various ratios of fast to slow rates, when the proportion of fast sites is fixed at 10%. \* represents the parameter values chosen in subsequent simulations.

Scenario 1 (overall fast rates and high rate ratio). By simulating sequences on the same tree topology and relative branch lengths, we can directly compare the inferred overall rate and age estimates between trees of 15 different depths (i.e., root ages). Results from 1 of the 10 tree topologies used in this study are presented in figures 4 and 6 as an illustration of the time-dependent patterns. Results from all 10 topologies are compared in supplementary material, Supplementary Material online.

### Rates Estimated with and without a Gamma Model

As previously shown by the mathematical analysis, when the model does not correct for among-site rate heterogeneity



**FIG. 4.** Relationship between simulated and inferred substitution rates and node ages. (A) Rate estimates reported for 15 trees of varying root age, and their relationship to the real rates. The pink curve represents the rates calculated mathematically using the same rates, rate ratio, and proportion of fast sites as those used for simulation. (B) Ratio of the estimated to expected (simulated) node ages for the node placed at 5% of the root age on the simulated tree. For both graphs, the empty shapes mark the absence of admixture (nonconvergence) for at least one parameter of the Bayesian analysis.

(no gamma), the overall rates inferred from the simulated alignment are underestimated in a time-dependent pattern (red curve in [fig. 4A](#)). In that situation, the correct rate is recovered only for the shallowest tree, with a root age of 1 ky (“correct” means the simulated rate was included within the 95% highest posterior density interval of the inferred rate). To show the correlation between the theoretical results and the phylogenetic inferences, the curve calculated mathematically with the same parameters is plotted on the same graph (pink curve in [fig. 4A](#)).

When a gamma distribution is used to model the RHAS, the correct rate is recovered for only 4 of the 15 trees, but the pattern of the bias is completely different (blue curve in [fig. 4A](#)). For the four shallowest trees (root ages 1–10 ky), we observe a similar pattern to the noncorrected curve but with reduced underestimates, showing that the gamma model is improving rate inferences at this stage. When the root age is 20 ky, this trend is reversed and the curve turns upwards, with large rate overestimates until the root age reaches 2 My. A second inversion appears at 5 My, leading to rate underestimates in the four deepest trees (root ages >5 My). For the two deepest trees (root ages 20 and 50 My), the BEAST analyses did not converge (effective sample sizes <100, empty shapes in [fig. 4A](#)); at the simulated rate of evolution, these sequences are completely saturated (see text below and [fig. 5](#)). They are presented on the graphs to show the limit of analyzable alignments, delimiting the time range for which phylogenetic signal is present in the simulated sequences.

### Saturation

Saturation is the principal cause of degradation of the phylogenetic signal through time. To investigate more thoroughly the presence of signal in the sequences, the saturation level for each of the alignments generated on the 15 simulated trees is depicted in [figure 5](#) (and [supplementary figure 10](#), [Supplementary Material](#) online). Three examples of saturation plots from the concatenated alignments are shown in [figure 5A](#). At a root age of 5 ky, there is almost no sign of saturation, whereas at a root age of 0.5 My (i.e., 500 ky), a low level of saturation starts to appear (i.e., an observed distance of 15% can correspond to an inferred genetic distance of 20–30%). The sequences are completely saturated at a root age of 50 My.

The saturation levels at all root ages are compiled in [figure 5B](#), either for the concatenated alignment (dashed line) or both subsets individually (plain lines). This last figure shows the importance of the differential saturation between the fast and slow subsets of the alignment through time. The fast sites become completely saturated from 50 ky, whereas the slow sites start to present signs of saturation ~1 My, to reach full saturation at 50 My. Consequently, the concatenated sequences present an intermediate saturation pattern, approaching a plateau between 50 ky and 2 My. Interestingly, the time frame of this plateau, corresponding to the highest degree of differential saturation between fast and slow sites, matches the time frame of strong rate

overestimation when a gamma distribution is used (blue curve in [fig. 4A](#)). Furthermore, and importantly, this maximum differential saturation matches the time frame over which there were greatest errors in node age estimates, driven by errors in estimating relative branch lengths, as discussed later.

For two trees (root ages 1 and 2 My), when the gamma distribution was used, both the gamma shape parameter and the mean rate failed to converge (effective sample sizes <10, empty shapes on [fig. 4A](#)). It is important to note that this convergence issue is not observed with PhyTime (see text below and [supplementary fig. 6](#), [Supplementary Material](#) online). As all the other parameters of the analysis converged satisfactorily, it appears that the difficulty in estimating the shape parameter of the gamma distribution is related to the observed convergence issue. Each change in the pattern of the rate estimates corresponds to a change in the inferred shape parameter ([supplementary fig. 9](#), [Supplementary Material](#) online). This suggests that the ability of the program to estimate correctly the shape of the gamma distribution is related to the observed pattern of time-dependent rate estimates.

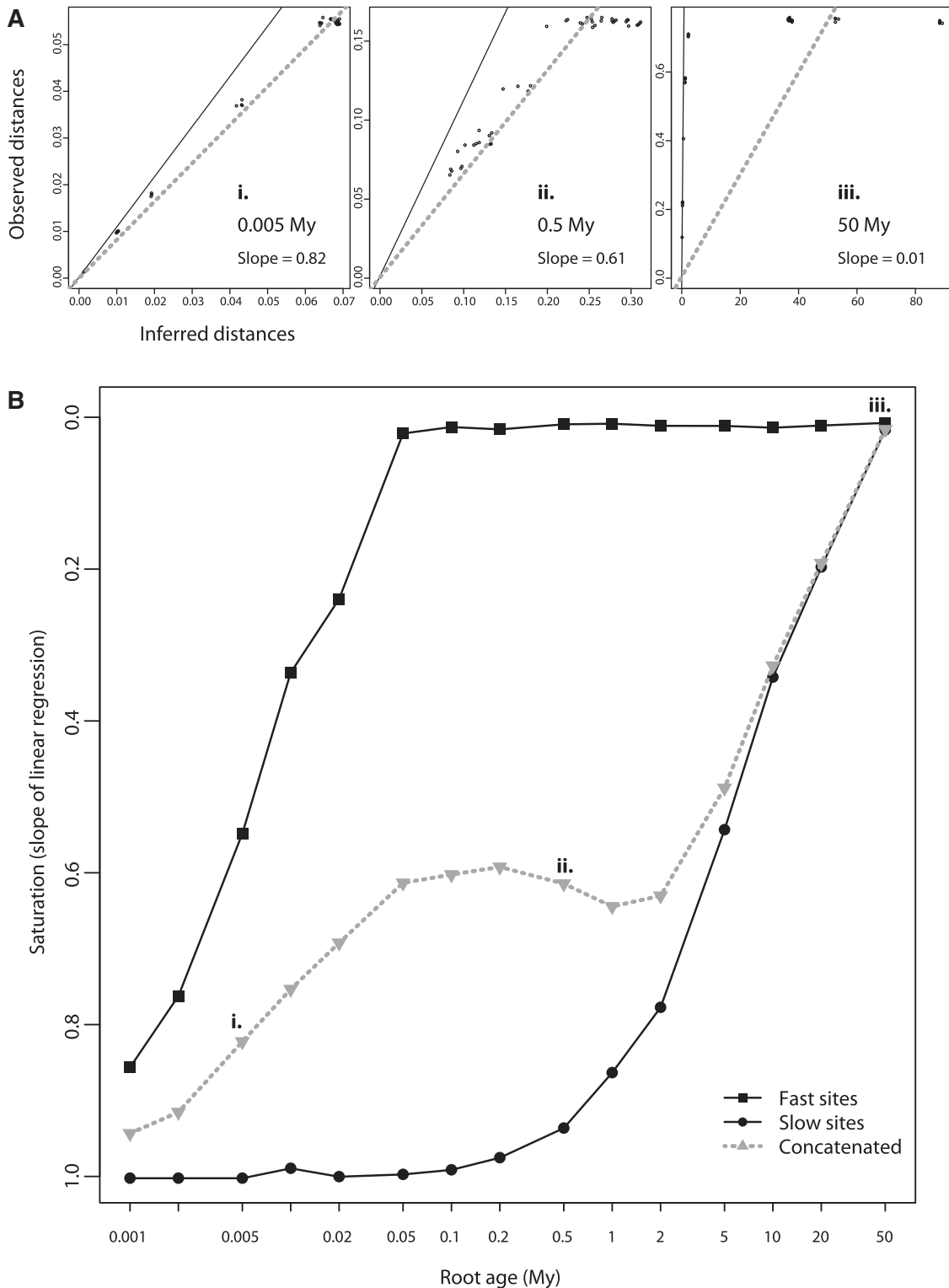
### Consequences for Molecular Estimates of Divergence Times

The primary aim of molecular-clock analysis is often the dating of divergences in a tree, rather than estimating the average substitution rate. When the ages of calibrating nodes are known accurately, as in these analyses, accurate estimation of relative branch lengths will be sufficient for accurate divergence dating. Therefore, to observe the consequences of the time-dependent rate estimates on relative node heights, the inferred ages of the nodes are contrasted with the actual (simulated) dates.

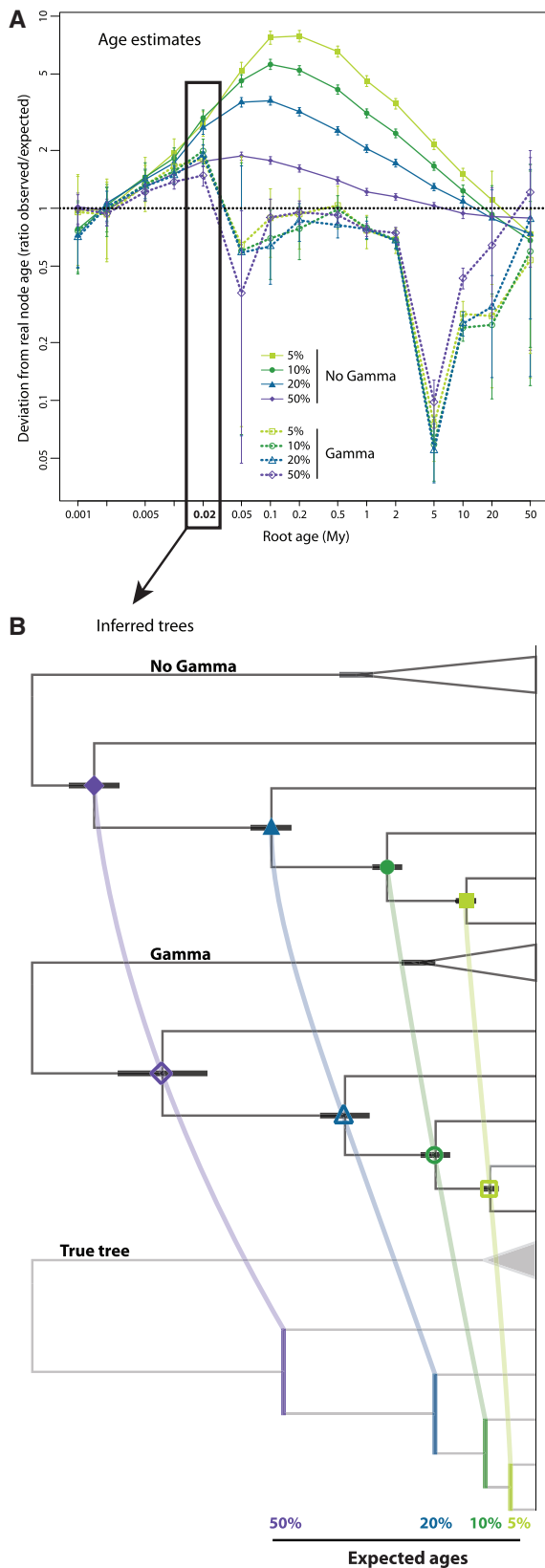
The results for the node initially set at 5% of root age are presented first as an example in [figure 4B](#). When among-site rate variation is not modeled (no gamma, red curve), the correct node age is retrieved for only 4 of the 15 tree depths. For shallower trees (root age 1 ky to 0.2 My) the 5% node age is gradually overestimated, reaching 7.9 times its expected value at 0.2 My; this node is expected to be 10 ky old, but has an inferred age of 73–84 ky. Because the correct age of the 100% (root) node is always used as a calibration, the overestimate of the age of the 5% node cannot be simply due to a general underestimation of evolutionary rate across the entire tree. Rather, it demonstrates that the relative branch lengths in the inferred tree are distorted, with basal branches being shortened in comparison to the branches near the tips (see below).

Surprisingly, for trees with root ages of 0.5 to 50 My, the bias in age estimation is reduced. Although overall rates are increasingly underestimated, the age of the 5% node is more accurately inferred. The correct age for the 5% node is estimated for the two shallowest trees, which show the least saturation (1 and 2 ky), and for the two deepest trees, which show the most saturation (20 and 50 My) but for which the differential saturation is at a minimum ([fig. 5B](#)). The 5% node described here presents the biggest bias, because it is furthest from the calibration point (root), but all four





**Fig. 5.** Saturation estimates from the simulated sequences. (A) The level of saturation was estimated by comparing the genetic distances inferred from a ML tree (X axis) to the pairwise distances directly observed from the sequences (Y axis). To quantify the phenomenon, we use a linear regression through the origin (dashed line): the lower the slope, the more saturated the sequences. The diagonal (black line) represents the ideal “no saturation” state. Note that both axes vary in scale across the three plots. (i) At 5 ky, there is almost no sign of saturation, whereas (ii) at 0.5 My, we start to observe a low level of saturation (i.e., an observed distance of 0.15 can correspond from 0.20 to 0.30 inferred genetic distance). (iii) At 50 My, the sequences are completely saturated. (B) The slope values, representing the degree of saturation of the sequences, are reported for the concatenated alignments (dashed line) and for both fast and slow subsets (plain lines).



**FIG. 6.** Distortion of retrieved node ages for four simulated nodes (5, 10, 20, and 50% of root/calibration age). (A) Ratio of the inferred to expected (simulated) node ages, for trees of 15 different root ages. (B) Comparison of inferred trees without (upper) and with (center) gamma to the actual tree (lower), for simulated tree with root age of 20 ky. Note the artifactually shortened basal branches in both inferred trees, with the bias being greater in the upper tree.

nodes, fixed in the tree used for simulation, show the same pattern (fig. 6A): for shallower trees, all branches are less saturated, and for deeper trees, all branches are highly saturated. In the latter case, even though the overall evolutionary rate is underestimated, relative branch lengths across the tree are not greatly distorted. At intermediate ages, the largest dating errors occur, due to the strongest differential saturation across branches: the recent branches are far less saturated than the deeper branches, resulting in greatly distorted relative branch lengths.

When a gamma distribution is used to model RHAS, the correct node age is inferred for 7 of the 15 trees (blue curve in fig. 4B). Between 1 and 20 ky, the estimates follow the tendency described earlier when no gamma distribution was used, showing a gradual overestimation of the node ages. Between 50 ky and 0.5 My, the correct value is inferred (with large uncertainty for the 50 ky calibration), despite substantial overestimation of the rates. For trees of these ages, the use of a gamma distribution appears to be efficiently recovering the correct tree proportions even in this extreme situation of among-site rate heterogeneity. For trees with root calibrations deeper than 1 My, the node ages are consistently underestimated. There is an extreme case at 5 My, where the 5% node is simulated to be 250 ky old but is inferred to be between 12 and 26 ky. Importantly, this is not one of the two analyses for which the Markov chain failed to converge for the rate and the shape parameter of the gamma distribution, potentially leading to false confidence in the pattern of bias. In agreement with the results from the overall rate estimates, trends in the estimation of node ages with increasing tree depth appear to change when the estimated shape parameter of the gamma distribution is changing abruptly.

We can compare the consequences of the time-dependent rate estimates on the estimates of node ages at different depths in the trees (5, 10, 20, and 50% of the basal node; respectively colored in orange, purple, blue, and green in fig. 6). When among-site rate heterogeneity is not modeled using a gamma distribution, the largest overestimation is observed for the younger nodes (fig. 6A) showing a correlation between the distance from the calibration point (here the root) and the magnitude of the bias. This observation is confirmed by the replication of these analyses using a calibration at the 5% node instead of the root, and estimating the ages of all other nodes including the root (supplementary fig. 8, Supplementary Material online). In this situation, the age estimates for the root of the tree (furthest from the new shallow calibration point) have the greatest bias. This correlation is not observed when a gamma distribution is used to model rate variation among sites. The phylogenetic estimate for the simulated tree with root age of 20 ky is presented in figure 6B. Although this relatively shallow tree does not show the most biased results, the impact on the tree proportions is obvious.

#### Confirmation of Results

The results from Scenario 2 are shown in supplementary figures 3 and 4, Supplementary Material online. This scenario presents lower overall rates, and a lower rate ratio between

slow and fast sites, than Scenario 1. The same time-dependent patterns described for Scenario 1 were retrieved for estimated rates and node ages in Scenario 2. However, the lower overall rates used to simulate the sequences leads to a predictable shift in the time taken for the biases to manifest: the rate is overestimated and the node ages underestimated after 200 ky, instead of 2 ky for Scenario 1. The effect of a lower rate ratio between slow and fast sites also has consequences on the estimates of rates and tree proportions, with the observed biases being smaller.

Similar time-dependent results were obtained using: multiple replicates, a Yule prior in BEAST, PhyTime and a three-rate scenario with intermediary molecular rate. This shows that the observed time dependence of inferred rates and node ages is neither due to one particular dataset, nor that it is specific to the program or method used to simulate the alignments. The results from the replicates are presented in the [supplementary material, Supplementary Material](#) online.

#### *Improvement Using a More General Mixture Model*

One solution for dealing with RHAS when the site-specific rate variation does not follow a gamma distribution is to use a more complex mixture model. The gamma distribution is a mathematically tractable but very restricted type of mixture model that assumes that sites vary only in overall substitution rate (and not other substitution parameters), that they fall into a specific number of rate categories, and that the distribution of site-specific rates fits a gamma distribution. More general mixture-models can relax these assumptions and also capture other variation, such as different substitution matrices and different base frequencies.

To evaluate the efficacy of such models in the presence of strong RHAS, we implemented a new Bayesian mixture model called “FreeRate” in the program PhyTime. The “FreeRate” option allows a pre-determined number of classes of substitution rates among sites, with the posterior distributions of the rates for each class and the corresponding proportions being estimated from the data. Crucially, the distribution of rates is not forced to follow the gamma (or any other parametric) distribution.

The results of analyses performed with the “FreeRate” option on our simulated alignments are compared with the results from BEAST in [figure 4](#) (orange curves). The correct overall rate is inferred for the seven shallowest trees, after which the rate is progressively underestimated ([fig. 4A](#)). Although the estimated rates are time dependent after 0.2 My in this situation, the FreeRate model outperformed the gamma-distribution model implemented by both BEAST ([fig. 4A](#)) and PhyTime ([supplementary fig. 6, Supplementary Material](#) online). More importantly, the correct node ages were inferred for 10 of the 15 tree depths (2 of the 5 incorrectly inferred values being for the fully saturated alignments at 20 and 50 My). This demonstrates the capability of a general mixture-model to recover the correct tree proportions from an alignment with substantial RHAS, where a more restricted mixture model (gamma distribution) appears to be inadequate.

*Improvement from Partitioning the Data According to RHAS*  
The use of a general mixture model is an effective way to estimate and characterize site-specific rate heterogeneity in an alignment, but one might also have a prior knowledge or expectation of such heterogeneity (e.g., codon positions in protein coding genes).

Using our simulated data, when the fast and slow sites are treated as two separate data subsets in BEAST (green curves in [fig. 4](#)), the fast rate is correctly inferred for eight tree depths (two additional estimates being very close to the correct rate) up to 1 My. At deeper tree depths, the upper ends of the 95% highest posterior density intervals of these rate estimates approach infinity, in agreement with the complete saturation observed for these fast-evolving sites ([fig. 5B](#)). For the slow rate, BEAST estimated the correct value for 11 of the 15 tree depths, none of the estimates showing substantial deviation from the expected rate. More importantly, from a dating perspective, the correct age for the 5% node was recovered for 13 of these 15 tree depths ([fig. 4B](#)). This analysis demonstrates the presence of recoverable phylogenetic signal through the 15 tree depths of our study. It also shows that prior knowledge of the location of mutational hotspots along sequences can be used to improve the estimation of the overall substitution rate and the node age estimates.

## Conclusion

Our study confirms that RHAS can lead to a time-dependent pattern of rate estimates. We have presented a mathematical characterization of the phenomenon, and using simulations we have explored the consequences on the estimated rates and node ages when conducting phylogenetic analyses. Our simulations show that the greatest bias in age estimates occurs at intermediate (rather than the oldest) root ages. Relative branch lengths are least distorted in very shallow trees (where mutational saturation is absent), and in very deep trees (where all branches are highly and uniformly saturated, and to roughly the same extent). At intermediate ages, where only basal branches are saturated (and thus shortened), relative branch lengths are most distorted, leading to maximal bias in age estimates.

Our results show that assuming a distribution-free mixture model (Pagel and Meade 2004; Huelsenbeck and Suchard 2007), or simply a gamma mixture model (Mayrose et al. 2005) would limit these biases. Also, we demonstrate that an a priori identification of the fastest sites of a particular alignment using existing tools, such as IQPNNI (Meyer and von Haeseler 2003) or mixture model algorithms, would allow the partitioning of the data in BEAST and improve the estimates of rates and node ages (providing sites are assigned to the adequate rate classes).

Although the observed time dependence of molecular rates is not solely explained by RHAS, the presence of strong rate heterogeneity in the data clearly affects both the estimation of the node age and substitution rate. This further emphasizes the need for the development and use of appropriate models of molecular evolution.

**Supplementary Material**

Supplementary material is available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

**Acknowledgments**

The authors thank Alexei Drummond for providing helpful comments on the research design, and Bastien Llamas for help with figure development and helpful comments on the manuscript. They thank Jack Da Silva, Wolfgang Haak, David Penny, Peter Lockhart, Paul Muscat, Craig Mitchell, and Laura French for their participation in the early stages of the research. They also thank three anonymous reviewers for helpful comments on the manuscript. J.S., C.D.S., S.Y.W.H., and A.C. were supported by the Australian Research Council, and S.G. is supported by the Royal Society of New Zealand through a Marsden grant.

**Appendix: Technical Details and Proof of Theorem 1**

We first outline how the earlier analysis extends to accommodate continuous distributions of rates across sites in the correction model *M*. In this case, equation (6) is replaced by

$$f(x) = \int_0^\infty \gamma [1 - \exp(-xr/\gamma)] \lambda(r) dr,$$

where  $\lambda(r)$  is the probability density function for the continuous distribution of rates, scaled so that the mean rate is 1 (i.e.,  $\int_0^\infty r\lambda(r)dr = 1$ ); this is applied as before in equations (5) and (7). Note that we can write  $f(x) = \gamma [1 - \psi_M(-x/\gamma)]$ , where  $\psi_M$  is the moment generating function of the distribution of rates across sites. For example, for a gamma distribution with shape parameter *k* we have  $\psi_M(x) = (1 - x/k)^{-k}$  and so

$$f(x) = \gamma [1 - (1 + \frac{x}{k\gamma})^{-k}]. \tag{8}$$

These formulae can be extended further to allow *M* to be a mixture of discrete and continuous rate distributions (e.g., in the case of a continuous distribution for variable sites, plus a discrete class class of invariable sites).

Next, we describe how to apply equation (6) when the correction model has just one rate class or is a gamma distribution. For a single rate class, we saw earlier (equation 7) that  $r(t) = \frac{g'(t)}{1-g(t)/\gamma}$ . Thus, for a discrete mixture of rates ( $r_i$  with frequency  $a_i$ ) in the generating model (G) we have

$$\hat{r}(t) = \mu \frac{\sum_{i=1}^n \alpha_i r_i \exp(-\mu r_i t / \gamma)}{[\sum_{i=1}^n \alpha_i \exp(-\mu r_i t / \gamma)]} \tag{9}$$

More generally, if the correction model is based on a (continuous) gamma distribution with shape parameter *k* then from equations (8) and (6) we obtain

$$\hat{r}(t) = \mu \frac{\sum_{i=1}^n \alpha_i r_i \exp(-\mu r_i t / \gamma)}{[\sum_{i=1}^n \alpha_i \exp(-\mu r_i t / \gamma)]^{1+k}} \tag{10}$$

Notice that, in the limit as  $k \rightarrow \infty$ , the gamma distribution approaches the single-rate distribution, and, accordingly, equation (10) converges to equation (9).

**Proof of Theorem 1**

For part (a) routine calculus shows that

$$\frac{d\hat{r}(t)}{dt} = \frac{g''(t)f'(\hat{k}_t) - g'(t)\frac{d}{dt}f'(\hat{k}_t)}{f'(\hat{k}_t)^2}$$

and as

$$\frac{d}{dt}f'(\hat{k}_t) = f''(\hat{k}_t)\frac{d}{dt}\hat{k}_t = f''(\hat{k}_t)\hat{r}(t) = f''(\hat{k}_t)g'(t)/f'(\hat{k}_t),$$

(by eqs. 1 and 6), we obtain

$$\frac{d\hat{r}(t)}{dt} = \frac{g''(t)f'(\hat{k}_t) - f''(\hat{k}_t)[g'(t)]^2/f'(\hat{k}_t)}{f'(\hat{k}_t)^2}.$$

Now, in the limit as *t* approaches 0,  $g'(t)$  approaches  $\mu$  and  $g''(t)$  approaches  $\frac{-\mu^2}{\gamma} \sum_{i=1}^n \alpha_i r_i^2$ , whereas  $f'(\hat{k}_t)$  approaches 1 and  $f''(\hat{k}_t)$  converges to  $\frac{-1}{\gamma} \sum_{i=1}^{n'} \alpha'_i (r'_i)^2$ . Thus

$$\frac{d\hat{r}(t)}{dt} = \frac{-\mu^2}{\gamma} [\sum_{i=1}^n \alpha_i r_i^2 - \sum_{i=1}^{n'} \alpha'_i (r'_i)^2] = \frac{-\mu^2}{\gamma} (\sigma_G^2 - \sigma_M^2),$$

where  $\sigma_G^2$  and  $\sigma_M^2$  is the variance of the generating process and correction model, respectively.

For Part (b-i), note that the  $r_i$ 's are ordered in increasing values (and the  $\alpha_i$  values are strictly positive) and so equation (9) implies that, provided  $r_1 > 0$ ,  $\hat{r}(t)$  is asymptotic to  $\mu\alpha_1 r_1 / \alpha_1 = \mu r_1$ . On the other hand, if  $r_1 = 0$ , then the denominator in equation (9) has a constant term ( $\alpha_1$ ) while all the numerator terms decay exponentially, and so  $\hat{r}(t)$  converges to zero. In either case  $\hat{r}(t)$  converges to  $\mu r_1$  (the smallest absolute rate of *G*).

For part (b-ii), observe that as  $t \rightarrow \infty$  then  $\hat{k}_t \rightarrow \infty$  and that, as  $r_1, r'_1 > 0$  we can write

$$1 - \frac{f(\hat{k}_t)}{\gamma} = \alpha'_1 \exp(-\hat{k}_t r'_1 / \gamma) [1 + h_1(\hat{k}_t)]; \tag{11}$$

$$1 - \frac{g(t)}{\gamma} = \alpha_1 \exp(-\mu r_1 / \gamma) [1 + h_2(t)],$$

where  $h_1(x)$  and  $h_2(x)$  is a sum of positively weighted exponential terms, which all decay to zero. Rewriting equation (5) as  $1 - \frac{f(\hat{k}_t)}{\gamma} = 1 - \frac{g(t)}{\gamma}$  and taking the logarithm of both sides of this equation using equation (11) gives

$$\log(\alpha'_1) - \hat{k}_t r'_1 / \gamma + \log[1 + h_1(\hat{k}_t)] = \log(\alpha_1) - \mu r_1 / \gamma + \log[1 + h_2(t)],$$

and differentiating both sides with respect to  $t$  (and observing that  $d\hat{\kappa}_t/dt = \hat{r}(t)$ ) gives

$$-\hat{r}(t)r'_1/\gamma + \frac{h'_1(\hat{\kappa}_t)r(t)}{1+h_1(\hat{\kappa}_t)} = -\mu r_1/\gamma + \frac{h'_2(t)}{1+h_2(t)},$$

which gives

$$\hat{r}(t) = \mu r_1/r'_1 - b(t) + a(t)r(t), \quad (12)$$

where  $a(t) = \frac{\gamma h'_1(\hat{\kappa}_t)}{r'_1[1+h_1(\hat{\kappa}_t)]}$  and  $b(t) = \frac{\gamma h'_2(t)}{r'_1[1+h_2(t)]}$ .

Rearranging equation (12) gives

$$\hat{r}(t) = \frac{\mu r_1/r'_1 - b(t)}{1 - a(t)}, \quad (13)$$

Now,  $b(t)$  clearly converges to 0 with increasing  $t$ , and so too does  $a(t)$  because  $\hat{\kappa}_t$  tends to infinity as  $t \rightarrow \infty$ . Thus, by equation (13), we deduce that  $\hat{r}(t)$  converges to  $\mu r_1/r'_1$ , as claimed.

For Part (b-iii), as  $r_1 > 0$ , the numerator of equation (10) is asymptotic to  $\mu \alpha_1 r_1 e^{-r_1 t/\gamma}$  as  $t \rightarrow \infty$  while the denominator is asymptotic to  $(\alpha_1 e^{-r_1 t/\gamma})^{1+\frac{1}{k}}$ . As  $k$  is finite, the denominator of equation (10) converges to 0 more quickly than the numerator, and so the rate curve tends to infinity.

For Part (b-iv) note that, as  $x$  increases,  $f(x)$  converges to  $\gamma(1 - i_M)$  where  $i_M$  is the proportion of invariable sites in  $M$ , and  $g(x)$  converges to  $\gamma(1 - i_G)$  where  $i_G$  is the proportion of invariable sites in  $G$ . Part (iv) now follows by considering the graphs of  $f(x)$  and  $g(x)$  for the cases  $i_M < i_G$  and  $i_M > i_G$  and thereby the constraints on the solution for  $\kappa_t$  in equation (5).

## References

- Baloux F, Lehmann L. 2011. Substitution rates at neutral genes depend on population size under fluctuating demography and overlapping generations. *Evolution* 66:605–611.
- Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *Plos Biol.* 4:699–710.
- Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.
- Galtier N, Enard D, Radondy Y, Bazin E, Belkhir K. 2006. Mutation hot spots in mammalian mitochondrial DNA. *Genome Res.* 16:215–222.
- Gignoux CR, Henn BM, Mountain JL. 2011. Rapid, global demographic expansions after the origins of agriculture. *Proc Natl Acad Sci U S A.* 108:6044–6049.
- Guindon S. 2010. Bayesian estimation of divergence times from large sequence alignments. *Mol Biol Evol.* 27:1768–1781.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.
- Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22: 160–174.
- Henn BM, Gignoux CR, Feldman MW, Mountain JL. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol Biol Evol.* 26:217–230.
- Ho SYW, Endicott P. 2008. The crucial role of calibration in molecular date estimates for the peopling of the Americas. *Am J Hum Genet.* 83:142–146.
- Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Mol Ecol.* 20:3087–3101.
- Ho SYW, Larson G. 2006. Molecular clocks: when times are a-changin'. *Trends Genet.* 22:79–83.
- Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol.* 22:1561–1568.
- Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B. 2008. The effect of inappropriate calibration: three case studies in molecular ecology. *Plos One.* 3:e1615.
- Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. 2007. Evidence for time dependency of molecular rate estimates. *Syst Biol.* 56:515–522.
- Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Hermstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am J Hum Genet.* 72: 659–670.
- Huelsenbeck JP, Larget B, Swofford D. 2000. A compound poisson process for relaxing the molecular clock. *Genetics* 154:1879–1892.
- Huelsenbeck JP, Suchard MA. 2007. A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol.* 56: 975–987.
- Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genetics.* 22:225–231.
- Jukes TH, Cantor CR. 1969. Evolution of protein molecules. New York (NY): Academic Press.
- Kjer KM, Honeycutt RL. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol Biol.* 7:8.
- Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet.* 6:654–662.
- Lambert DM, Ritchie PA, Millar CD, Holland B, Drummond AJ, Baroni C. 2002. Rates of evolution in ancient DNA from Adelie penguins. *Science* 295:2270–2273.
- Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.
- Mayrose I, Friedman N, Pupko T. 2005. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* 21: 151–158.
- Meyer S, von Haeseler A. 2003. Identifying site-specific substitution rates. *Mol Biol Evol.* 20:182–189.
- Page M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol.* 53:571–581.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
- Parsons TJ, Muniec DS, Sullivan K, et al. 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet.* 15:363–368.
- Philippe H, Snell EA, Baptiste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol.* 21:1740–1752.
- Pulquerio MJ, Nichols RA. 2007. Dates from the molecular clock: how wrong can we be? *Trends Ecol Evol.* 22:180–184.

- R Development Core Team. 2011. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
- Rambaut A, Drummond AJ. 2007. Tracer v1.4. [cited Jun 2011]. Available from: [beast.bio.ed.ac.uk/Tracer](http://beast.bio.ed.ac.uk/Tracer).
- Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*. 13:235–238.
- Rosset S, Wells RS, Soria-Hernanz DF, Tyler-Smith C, Royyuru AK, Behar DM. 2008. Maximum-likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics* 180:1511–1524.
- Semple C, Steel M. 2003. Phylogenetics. Oxford: Oxford University Press.
- Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT. 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Annu Rev Ecol Evol System*. 37: 545–579.
- Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*. 84:740–759.
- Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic Entomol*. 35:429–448.
- Subramanian S, Denver DR, Millar CD, Heupink T, Aschrafi A, Emslie SD, Baroni C, Lambert DM. 2009. High mitogenomic evolutionary rates and time dependency. *Trends Genet*. 25:482–486.
- Swofford DL. 2003. PAUP\*: phylogenetic analysis using parsimony, version 4.0b10.
- Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*. 15:1647–1657.
- Woodhams M. 2006. Can deleterious mutations explain the time dependency of molecular rate estimates? *Mol Biol Evol*. 23: 2271–2273.
- Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306–314.
- Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*. 11:367–372.
- Zuckerkandl E, Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press. p. 189–225.