

# Relaxed phylogenetics and dating with confidence



# Relaxed Phylogenetics and Dating with Confidence

**Alexei J. Drummond[☯¤], Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut[☯*]**

Department of Zoology, University of Oxford, Oxford, United Kingdom

**In phylogenetics, the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum. Despite their dominance in phylogenetic inference, it is evident that both are biologically unrealistic and that the real evolutionary process lies between these two extremes. Fortunately, intermediate models employing relaxed molecular clocks have been described. These models open the gate to a new field of "relaxed phylogenetics." Here we introduce a new approach to performing relaxed phylogenetic analysis. We describe how it can be used to estimate phylogenies and divergence times in the face of uncertainty in evolutionary rates and calibration times. Our approach also provides a means for measuring the clocklikeness of datasets and comparing this measure between different genes and phylogenies. We find no significant rate autocorrelation among branches in three large datasets, suggesting that autocorrelated models are not necessarily suitable for these data. In addition, we place these datasets on the continuum of clocklikeness between a strict molecular clock and the alternative unrooted extreme. Finally, we present analyses of 102 bacterial, 106 yeast, 61 plant, 99 metazoan, and 500 primate alignments. From these we conclude that our method is phylogenetically more accurate and precise than the traditional unrooted model while adding the ability to infer a timescale to evolution.**

Citation: Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4(5): e88. DOI: 10.1371/journal.pbio.0040088

## Introduction

From obscure beginnings, phylogenetics has become an essential tool for understanding molecular sequence variation. In the past decade, huge progress has been made in developing methods for inferring phylogenies and estimating divergence dates. This development has been characterized by increases, both in the complexity of the models used to describe molecular sequence evolution, and in the sophistication of the methods for analyzing these new models. Nevertheless, a well-known problem that has persistently troubled phylogenetic inference is that of substitution rate variation among lineages. In order to infer divergence dates, it is convenient to assume a constant rate of evolution throughout the tree [1,2]. This practice has been regularly challenged by results from datasets showing considerable departures from clocklike evolution [3–5], and rate variation among lineages can seriously mislead not only divergence date estimation [6] but also phylogenetic inference (e.g., [7,8]).

Such problems with the molecular clock hypothesis have resulted in it being abandoned almost entirely for phylogenetic inference in favor of a model that assumes that every branch has an independent rate of molecular evolution. Under such an assumption, it is possible to infer phylogenies (e.g., [9,10]), but not to estimate molecular rates or divergence times, because the individual contributions of rate and time to molecular evolution cannot be separated. If the rate and time along each branch can only be estimated as their product, then the position of the root of the tree cannot be estimated without additional assumptions such as an outgroup or a non-reversible substitution process. This unrooted alternative to the molecular clock was first suggested by Felsenstein [10] and has formed the basis of all modern phylogenetic inference and is implemented in all major phylogenetic packages (e.g., PHYLIP [11], PAUP* [12], and MrBayes [9]).

Recently, it has been realized that less drastic alternatives to the unrooted model of phylogeny may exist. Instead of dispensing with the molecular clock entirely, attempts have been made to relax the molecular clock assumption by allowing the rate to vary across the tree [13–15]. For example, local molecular clock models estimate a separate molecular rate for each user-circumscribed group of branches in the tree [6,13,16]. However, assigning branches to different groups can be a difficult exercise if the number of sequences is large or if there is considerable uncertainty about the phylogenetic relationships among the taxa. Essentially, such models are only useful in cases in which there is a strong prior hypothesis that the rate of specific taxa will differ from the rest of the tree [6].

Bayesian relaxed-clock methods, including those published by Thorne et al. [15] and Aris-Brosou and Yang [17], present an enticing alternative to local clock models. These model the

**Academic Editor:** David Penny, Massey University, New Zealand

**Received** May 16, 2005; **Accepted** January 23, 2006; **Published** March 14, 2006

**DOI:** 10.1371/journal.pbio.0040088



**Abbreviations:** ACED, autocorrelated exponential distribution; ACLN, autocorrelated lognormal distribution; bp, base pairs; CLOC, strict molecular clock; HPD, highest posterior density; MAP, maximum a posteriori; MCMC, Markov chain Monte Carlo; UCED, uncorrelated exponential distribution; UCLN, uncorrelated lognormal distribution; UF, unrooted Felsenstein

* To whom correspondence should be addressed. E-mail: andrew.rambaut@zoo.ox.ac.uk

☯ These authors contributed equally to this work.

¤ Current address: Department of Computer Science, University of Auckland, Auckland, New Zealand

molecular rate among lineages as varying in an autocorrelated manner, with the rate in each branch being drawn (a priori) from a parametric distribution whose mean is a function of the rate on the parent branch. For example, a lognormal distribution can be employed with the variance scaled relative to the length of the branch in units of time, implying that the evolutionary rate changes continuously along the branch. Alternatively, the use of an exponential distribution would imply that changes occurred at the nodes, with the size of the change being independent of the branch length.

Autocorrelation of rates from ancestral to descendant lineages will occur whenever the largest component of rate variation is due to inherited factors, whether these are life-history traits or biochemical mechanisms. As one looks over smaller and smaller timescales, the differences in such inherited factors become smaller relative to the variance caused by stochastic and uninherited factors (such as environmental or chance events). An alternative way of considering this is that the autocorrelation is so strong that very little of the variation in rate can be attributed to inherited factors. At the other extreme, over very long timescales, we might expect so much variation in the inherited determinants of rate that the autocorrelation from lineage to lineage begins to break down, especially with sparse taxon sampling. However, it is difficult to predict where the boundaries between these effects are and thus to specify what the degree of autocorrelation will be.

Relaxed-clock models present a potentially useful method for removing the assumption of a strict molecular clock, but a major shortcoming of the methods that have been proposed thus far is that they require the user to specify the tree topology. This is a problem because in many cases, important parts of the tree may be uncertain or unresolved, resulting in a number of plausible tree topologies. Furthermore, a molecular clock may have been assumed when estimating the input tree (for example to find a root), but rate variation among lineages can adversely affect phylogenetic inference (e.g., [7,8]). In some settings, the tree topology may actually be a nuisance parameter and some other aspect of the model (such as the variance in evolutionary rate, the effective population size, or the age of the most recent common ancestor) is the object of interest. Lastly, the assumption of a relaxed clock will alter the posterior probabilities of alternative tree topologies, so that the best tree under a relaxed-clock model may differ from the best tree under an unrooted or strict molecular clock model. For these reasons, a "relaxed phylogenetics" approach, in which the phylogeny and the divergence dates are co-estimated under a relaxed molecular clock, is preferred [18].

Here we present a Bayesian Markov chain Monte Carlo (MCMC) [19,20] method for performing relaxed phylogenetics that is able to co-estimate phylogeny and divergence times under a new class of relaxed-clock models. Its utility is demonstrated through simulation and on 871 real datasets. When absolute rates and divergence dates are estimated, we use probabilistic calibration priors, rather than point calibrations, since these more appropriately incorporate calibration uncertainties. We have implemented this method in the application BEAST [21] in which they can be used in conjunction with a wide range of other evolutionary models.

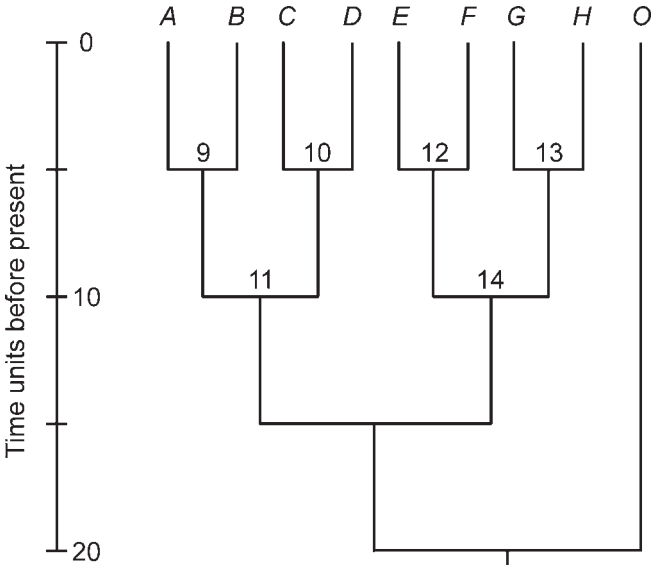


**Figure 1.** The Rooted Binary Tree Used for Simulating Sequence Evolution

The timescale is drawn in arbitrary time units. Apart from the branch leading to the outgroup, sequence O, all branches are five time units in length.

DOI: 10.1371/journal.pbio.0040088.g001

## Results

### Simulations

We generated alignments of nine nucleotide sequences, each 1,000 nucleotides in length, on the rooted tree in Figure 1. The outgroup sequence, O, is only used for rooting; otherwise, the tree is symmetric. Simulations were performed using the program RateEvolver v1.0 [22], which can simulate nucleotide substitution under different rate conditions, including constant (molecular clock), autocorrelated, and uncorrelated rates.

Fifty sequence alignments were generated under each of five sets of rate variation models: (1) Rates were fixed at 0.01 average substitutions per site per time unit throughout the tree (i.e., rates conformed to a molecular clock) (CLOC); (2) rates were lognormally autocorrelated among branches, with an ancestral rate of 0.01 average substitutions per site per time unit and a variance parameter ($S^2$) of 0.1, so that $S^2t = 0.5$ (ACLN); (3) rates were exponentially autocorrelated among branches, with an ancestral rate of 0.01 average substitutions per site (ACED); (4) rates were uncorrelated, with the rate in each branch independently drawn from a lognormal distribution with mean 0.01 and variance parameter of 0.5 (UCLN); and (5) rates were uncorrelated, with the rate in each branch independently drawn from an exponential distribution with mean (and therefore standard deviation) of 0.01 (UCED).

A normally distributed calibration prior with mean 20.0 and standard deviation 1.0 was specified for the age of the root of the tree, and the tree topology was fixed. Each alignment was analyzed using BEAST [21] with 5,000,000 steps, following a discarded burn-in of 500,000 steps. In each analysis, convergence of the chain to the stationary distribution was confirmed by inspection of the MCMC samples using the program Tracer 1.2 [23]. This application analyses posterior samples of continuous parameters from Bayesian

**Table 1.** The Proportion of Datasets (50 Simulations for Each of Five Models) for Which the True Rate Was within the 95% HPD Limits at the Given Branch

| Simulated Model | CLOC | | | ACLN | | | ACED | | | UCLN | | | UCED | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Analyzed Model** | **CLOC** | **UCED** | **UCLN** | **CLOC** | **UCED** | **UCLN** | **CLOC** | **UCED** | **UCLN** | **CLOC** | **UCED** | **UCLN** | **CLOC** | **UCED** | **UCLN** |
| Sequence A | 1.00 | 1.00 | 1.00 | 0.12 | 1.00 | 0.92 | 0.02 | 0.58 | 0.36 | 0.18 | 1.00 | 0.94 | 0.02 | 0.94 | 0.78 |
| Sequence B | 1.00 | 1.00 | 1.00 | 0.02 | 1.00 | 0.94 | 0.06 | 0.66 | 0.38 | 0.06 | 1.00 | 0.96 | 0.10 | 0.98 | 0.80 |
| Sequence C | 1.00 | 0.96 | 1.00 | 0.12 | 1.00 | 0.94 | 0.02 | 0.58 | 0.46 | 0.10 | 1.00 | 0.90 | 0.02 | 1.00 | 0.92 |
| Sequence D | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 0.92 | 0.02 | 0.62 | 0.52 | 0.10 | 1.00 | 0.90 | 0.16 | 0.94 | 0.84 |
| Sequence E | 1.00 | 1.00 | 1.00 | 0.08 | 0.96 | 0.90 | 0.00 | 0.62 | 0.50 | 0.06 | 1.00 | 0.90 | 0.10 | 0.94 | 0.82 |
| Sequence F | 1.00 | 1.00 | 1.00 | 0.08 | 0.98 | 0.90 | 0.06 | 0.70 | 0.38 | 0.10 | 1.00 | 0.96 | 0.04 | 0.96 | 0.74 |
| Sequence G | 1.00 | 1.00 | 1.00 | 0.12 | 0.98 | 0.88 | 0.02 | 0.72 | 0.56 | 0.20 | 1.00 | 0.92 | 0.10 | 0.96 | 0.80 |
| Sequence H | 1.00 | 1.00 | 1.00 | 0.06 | 1.00 | 0.92 | 0.06 | 0.72 | 0.50 | 0.12 | 1.00 | 0.90 | 0.04 | 0.96 | 0.80 |
| Node 09 | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 0.98 | 0.04 | 0.78 | 0.54 | 0.06 | 1.00 | 0.94 | 0.10 | 0.98 | 0.84 |
| Node 10 | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 1.00 | 0.00 | 0.76 | 0.62 | 0.14 | 1.00 | 0.96 | 0.18 | 1.00 | 0.80 |
| Node 11 | 1.00 | 0.98 | 1.00 | 0.08 | 1.00 | 1.00 | 0.04 | 0.90 | 0.68 | 0.12 | 1.00 | 0.92 | 0.06 | 0.96 | 0.84 |
| Node 12 | 1.00 | 1.00 | 1.00 | 0.10 | 0.98 | 0.98 | 0.02 | 0.80 | 0.62 | 0.16 | 1.00 | 0.94 | 0.08 | 0.98 | 0.82 |
| Node 13 | 1.00 | 1.00 | 1.00 | 0.08 | 1.00 | 1.00 | 0.02 | 0.84 | 0.78 | 0.08 | 1.00 | 0.92 | 0.18 | 0.96 | 0.82 |
| Node 14 | 1.00 | 1.00 | 1.00 | 0.06 | 1.00 | 0.98 | 0.10 | 0.90 | 0.76 | 0.08 | 1.00 | 1.00 | 0.14 | 0.96 | 0.88 |
| Overall | 1.00 | 1.00 | 1.00 | 0.08[a] | 0.99 | 0.95 | 0.03[a] | 0.73[a] | 0.55[a] | 0.11[a] | 1.00 | 0.93 | 0.09[a] | 0.97 | 0.82[a] |
| Average HPD size | 0.002 | 0.056 | 0.006 | 0.006 | 0.047 | 0.024 | 0.003 | 0.022 | 0.016 | 0.009 | 0.049 | 0.023 | 0.002 | 0.025 | 0.025 |

Each cell reports the results at the branch above the specified node for a particular combination of simulated and analyzed rate models. Node labels correspond to those given in Figure 1. The simulated models were CLOC, ACLN, ACED, UCLN, and UCED. The models used for inference were CLOC, UCED, and UCLN.

[a]The overall false-positive rate is significantly greater than 5% for these combinations of simulated and analyzed modes.

DOI: 10.1371/journal.pbio.0040088.t001

MCMCs to allow visual inspection of the chain behavior, estimating of the effective sample size of parameters and the plotting of marginal posterior densities. The effective sample size is the number of independent samples that would be the equivalent to the autocorrelated samples produced by the MCMC. This provides a measure of whether the chain has been run for an adequate length (for example, if the effective sample sizes of all continuous parameters are greater than 200).

In four of the five cases, the uncorrelated relaxed-clock approach to estimating rates performed well (Table 1). In all cases, the rate estimates made under the UCED had the largest 95% highest posterior densities (HPDs). This can be most clearly seen in the rates estimated from sequences generated under a molecular clock, with the average 95% HPD size under the UCED model exceeding that under the UCLN model by an order of magnitude.

When the sequences were simulated under a molecular clock, the 95% HPD interval of the posterior rate estimate almost always contained the true rate under all three analysis models (Table 1). For sequences simulated under any of the other models, CLOC did extremely poorly, with the true rates included in the 95% HPDs between only 3% and 11% of the time. Clock estimates of rates from data generated under exponential rate models (ACED and UCED) were poorer than those from data generated under lognormal rate models (ACLN and UCLN); this was expected, since the variance of the exponential distribution is larger than those of the lognormal distributions in our simulations.

For the data generated under lognormal models (ACLN and UCLN), both of the uncorrelated models (UCED and UCLN) performed well with respect to coverage, with the 95% HPD containing the true rate between 93% and 100% of the time for individual branches. However, for the UCED model this was at the expense of power, with the average size of the HPDs being twice as large as those for the UCLN model.

For data generated under UCED, the UCED model performed better than UCLN with both models giving the same average size of HPDs, but with the latter model including the true rates in the HPDs slightly less often (82%). Neither model performed as well when the data were generated under an ACED model, with the true rate in the 95% HPD between 36% and 90% of the time.

The accurate estimation of molecular rates is important because it has a direct impact on the estimation of branch lengths, which can in turn affect the inferred tree topology. Collectively, the results provide a strong recommendation against assuming a molecular clock when analyzing data that have not evolved under clocklike conditions, but the uncorrelated relaxed-clock models also perform well when the data are clocklike. The results favor the use of the UCLN model in that it has an accuracy comparable to the UCED model, but it results in considerably smaller HPDs. In particular, because the UCLN model has the variance of the lognormal distribution as a parameter, it can better accommodate data that are close to being clocklike. This is not contradicting the findings of a previous simulation-based study [22], which suggested that the autocorrelated exponential model outperformed the lognormal model in rate estimation, because the uncorrelated models presented here are fundamentally different from autocorrelated models. Moreover, the previous simulation study considered only the accuracy of the estimates, and not their precision.

## Dengue Virus Type 4 and Human Influenza A Virus

We selected two virus datasets that were matched in the number of sequences ($n = 69$) and the time span over which the data had been sampled (17 y). The first dataset was a previously published sequence alignment of the *E* gene of dengue-4 virus (1,485 base pairs [bp]) from Puerto Rico [24].

**Table 2.** Rate and Variance Estimates for Two Viral Datasets (Human Influenza A Virus and Dengue Type-4 Virus)

| Virus | Clock Model | Parameters | Coefficient of Variation ($\sigma_r$) | Mean Rate | External Rate | Internal Rate | Log (Marginal Posterior) | Log (Tree Likelihood) | Log (Coalescent Prior) | Population Size Scaled by Generation Length |
|---|---|---|---|---|---|---|---|---|---|---|
| Dengue-4 | Clock | 1 | — | 0.00098 | — | — | −3939.60 | −3709.95 | −229.65 | 11.10 |
| | Exponential | 1 | 0.99 | 0.00113 | 0.00138 | 0.00088 | −3901.47 | −3690.44 | −211.03 | 8.55 |
| | Lognormal | 2 | 0.39 | 0.00099 | 0.00103 | 0.00094 | −3927.62 | −3701.53 | −226.10 | 10.58 |
| Influenza A | Clock | 1 | — | 0.00505 | — | — | −4367.76 | −4202.48 | −165.28 | 4.30 |
| | Exponential | 1 | 0.99 | 0.00551 | 0.00598 | 0.00502 | −4310.65 | −4164.33 | −146.33 | 3.27 |
| | Lognormal | 2 | 0.51 | 0.00517 | 0.00518 | 0.00519 | −4331.27 | −4172.54 | −158.74 | 3.92 |

DOI: 10.1371/journal.pbio.0040088.t002

The second dataset was an alignment of hemagglutinin sequences from human influenza A virus selected to have a similar time frame (1981–1998; see Protocol S1 for details). In both of these datasets, each sequence in the alignment represents a consensus of the viral population within a single infected human host at the time of sampling. Therefore, both genealogies represent the ancestral relationships between the virus populations in a sample of 69 infected people spanning a 17-y period. These two viral datasets, particularly influenza A virus, are expected to exhibit the effects of natural selection, given the nature of their life histories [24–26].

Both datasets were analyzed under the strict molecular clock and the UCLN and UCED models. For all analyses the HKY (Hasegawa-Kishino-Yano) model of nucleotide substitution [27] was used with gamma-distributed rate heterogeneity among sites [28]. Calibration information for the rate of evolution stems from the fact that each sequence has a date of sampling associated with it [29]. A constant-population coalescent prior was assumed [30]. For each combination of data and model, two independent MCMC analyses were each run for 10,000,000 steps, resulting in acceptable mixing as determined by Tracer 1.2 [23]. Given adequate sampling, these two runs were combined to obtain an estimate of the posterior distribution. The resulting estimates for the overall rate of evolution and coefficients of variation from the six analyses are presented in Table 2.

The estimated coefficient of variation, $\sigma_r$, was 0.39 for the dengue virus dataset under the UCLN model. This compares with 0.51 for the influenza A dataset, suggesting that the dengue virus sequences are evolving in a more clocklike manner than the influenza virus sequences. Under the UCED model, both datasets produce an estimated $\sigma_r$ of 0.99. Under the exponential distribution $\sigma_r$ should be equal to 1.0 by definition, and the small discrepancy arises because of the discretization procedure described in Materials and Methods. In both datasets the estimated average rate was higher under the UCED model, especially in the internal branches. This elevated rate also corresponded with a lower estimate of the effective population size. Figure 2 shows a tree topology sampled from the posterior of the UCLN analysis of the influenza A virus dataset.

## Marsupials

In addition to the viral sequences, we analyzed a marsupial dataset. The alignment contained concatenated nuclear protein–coding genes (*APOB, RAG1, IRBP, vWF,* and *BRCA1;* 3,772 bp) for 17 marsupials and seven (outgroup) placental mammals, obtained from Amrine-Madsen et al. [31].

The extensions to BEAST for inferring divergence times, described here, are well suited to the marsupial dataset. It possesses some phylogenetic uncertainty, so it is more reasonable to integrate over the posterior distribution of topologies than to assume a single true topology. Furthermore, the dataset includes taxa that have evolved to substantially different sizes, life histories, and niches, which are all hypothesized predictors of molecular rate variation [32,33].

The early fossil record of marsupials [34] is poorly known. As a result, point calibrations that utilize the oldest fossils that mark divergences of one group from another are likely to be substantial underestimates, whereas simply defining wide calibration bounds can poorly represent our understanding of the fossil record. We selected prior probability distributions as calibration priors (Table 3) with the intention of providing realistic assessments of the uncertainty associated with the fossil record [35].

First, to ascertain the joint prior distribution on the nodes of interest, the four calibration points, the Yule prior, and the reciprocal monophyly constraints were analyzed without any sequence data. The combined results of two runs of 10,000,000 steps are given in Table 3.

In order to analyze the marsupial data, we assumed a general time-reversible [36] model of nucleotide substitution with gamma-distributed rate heterogeneity among sites [28] and a proportion of invariant sites. In addition, we assumed a UCLN model of rate variation among branches in the tree. A Yule prior on branching rates was employed and the reciprocal monophyly of the ingroup and outgroup was assumed a priori. Four independent MCMC analyses were each run for 10,000,000 steps, resulting in acceptable mixing as determined by Tracer 1.2 [23]. These four runs were combined to obtain an estimate of the posterior distribution (Table 3). The 95% credible set of the marsupial analysis included 12 unique tree topologies, and the maximum a posteriori (MAP) tree topology accounted for 0.32 of the total posterior probability. The estimated rate of the fastest branch in the MAP topology was 2.7 times faster than that of the slowest branch. The mean rate of evolution was 0.944 substitutions per site per billion years (95% HPD: 0.817–1.073). The birth rate of the Yule prior was estimated to be 0.0133 (95% HPD: 0.0035–0.0234).

There was a slight tendency toward a positive correlation in

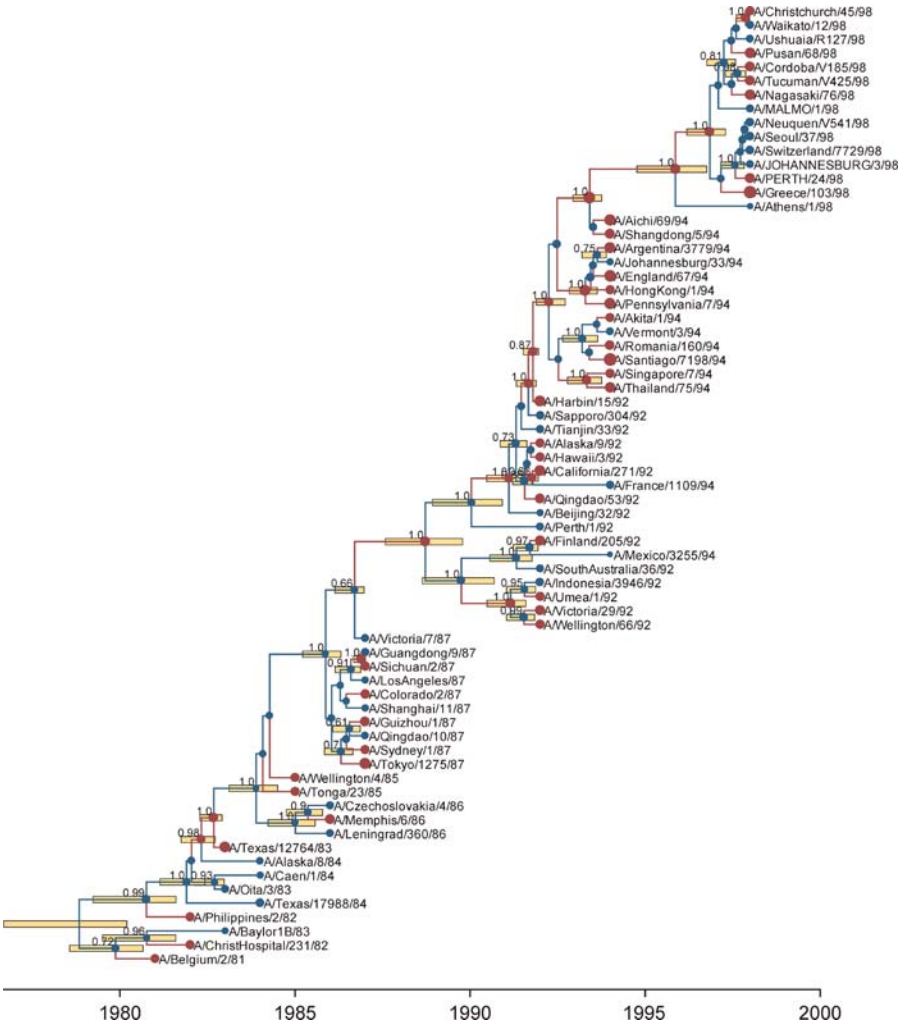


**Figure 2.** A Tree of 69 Influenza A Virus Sequences Drawn Randomly from the Posterior Distribution

The divergence times correspond to the mean posterior estimate of their age in years. The yellow bars represent the 95% HPD interval for the divergence time estimates. Both the mean and 95% HPD of the divergence times were calculated conditional on the existence of the clade defined by the divergence. Each node in the tree that has a posterior probability greater than 0.5 is labeled with its posterior probability. The sampling times of the tips were assumed to be known exactly. Branches colored in red had a posterior rate greater than the average rate, whereas branches colored in blue had a lower-than-average rate.

DOI: 10.1371/journal.pbio.0040088.g002

**Table 3.** Prior Probability Distributions and Posterior Probability Densities of the Marsupial Calibrations

| Calibration Node[a] | Prior | | MCMC Results | |
|---|---|---|---|---|
| | Distribution | Mean [95% CI][b] | Mean, No Data [95% HPD][c] | Mean, Posterior [95% HPD][d] |
| Elephants versus sirenians | Normal | 61.5 [52,71] | 61.3 [52.2,70.9] | 56.1 [46.2,65.2] |
| Dasyurids versus diprotodontians | Normal | 64 [54,74] | 64.2 [55.1,74.1] | 65.2 [56.7,73.7] |
| *Phascogale* versus *Dasyurus* | Normal | 17 [10,24] | 17.1 [10.5,24.3] | 14.4 [9.7,18.9] |
| Marsupials versus placentals | Translated Lognormal | 145 [132,180] | 148.4 [131.5,170.4] | 170.0 [140.0,204.6] |

[a]Calibration nodes are defined as the most recent common ancestor of the pair of taxa at any given step in the MCMC chain.
[b]The mean and 95% confidence intervals (CIs) of the prior probability distribution in millions of years.
[c]The mean and 95% HPD intervals of the posterior probability distribution in millions of years given by the MCMC procedure run without any sequence data. This will reveal the joint prior distribution on these parameters.
[d]The mean and 95% HPD intervals of the posterior probability distribution in millions of years.

DOI: 10.1371/journal.pbio.0040088.t003

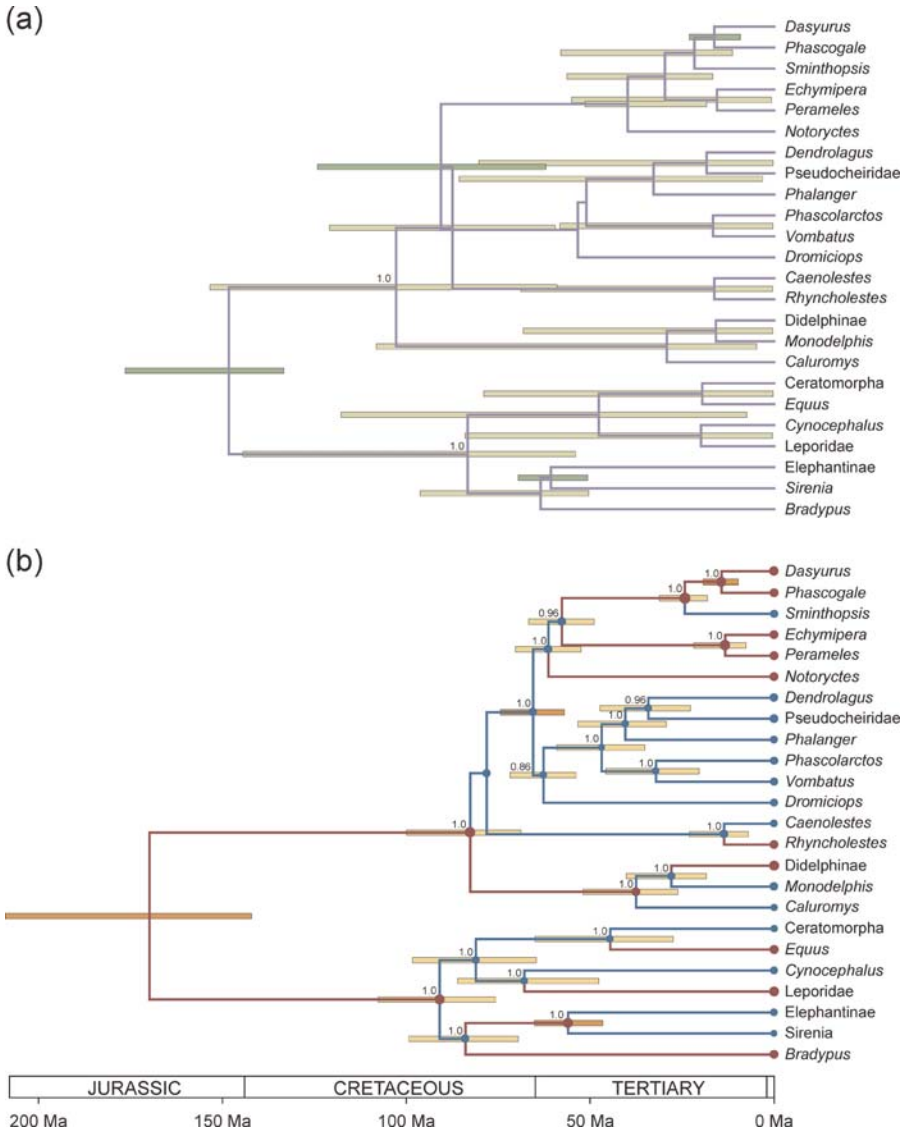


**Figure 3.** The Analysis of 17 Marsupials and Seven Placental Mammals

(A) The combined prior distribution of divergence times for the MAP tree topology. The green bars represent the 95% HPD interval for the divergence times. (B) The posterior distribution of the divergence times. The divergence times correspond to the mean posterior estimate of their age in millions of years. The yellow bars represent the 95% HPD interval for the divergence time estimates. Both the mean and 95% HPD of the divergence times were calculated conditional on the existence of the clade defined by the divergence. Each node in the tree is labeled with its posterior probability if it is greater than 0.5. The three nodes with normally distributed calibration priors are indicated by orange bars. Branches colored in red had a posterior rate greater than the average rate, whereas branches colored in blue had a lower-than-average rate.

DOI: 10.1371/journal.pbio.0040088.g003

the rate of parent and child branches but this was not significant (zero was included in the 95% HPD). The coefficient of variation was estimated to be 0.32 (95% HPD: 0.23–0.43), suggesting that the marsupial dataset is more clocklike than both of the virus datasets. Figure 3 shows the (A) prior and (B) posterior distributions of the clades present in the MAP tree topology.

## Autocorrelation of Rates amongst Lineages

We see no autocorrelation for the viruses we analyzed (the HPD interval of the covariance of parent and child branches was [–0.17,0.15] and [–0.18,0.15] for influenza and dengue-4 datasets respectively under the lognormally distributed model of rate variation and [–0.2,0.13] and [–0.19,0.13] for the exponentially distributed model of rate variation). For the marsupial dataset there is a small degree of autocorrelation suggested by the mean estimate, but it is not significantly different from zero (mean: 0.07, HPD: [–0.256, 0.4]). We would expect that larger datasets, particularly of diverse organisms that vary considerably in life-history traits or proofreading mechanisms, might exhibit substantial autocorrelation.

## Assessing Accuracy and Precision with Five Large Datasets

Five large datasets were obtained from previous studies: (1) amino acid alignments of 102 genes from eight bacterial species; (2) nucleotide alignments of 106 genes from eight yeast species [37]; (3) nucleotide alignments of 61 genes from nine plants; (4) amino acid alignments of 99 genes from nine metazoans; and (5) 500 nucleotide alignments of non-coding sequence from nine primates. The bacterial dataset was a

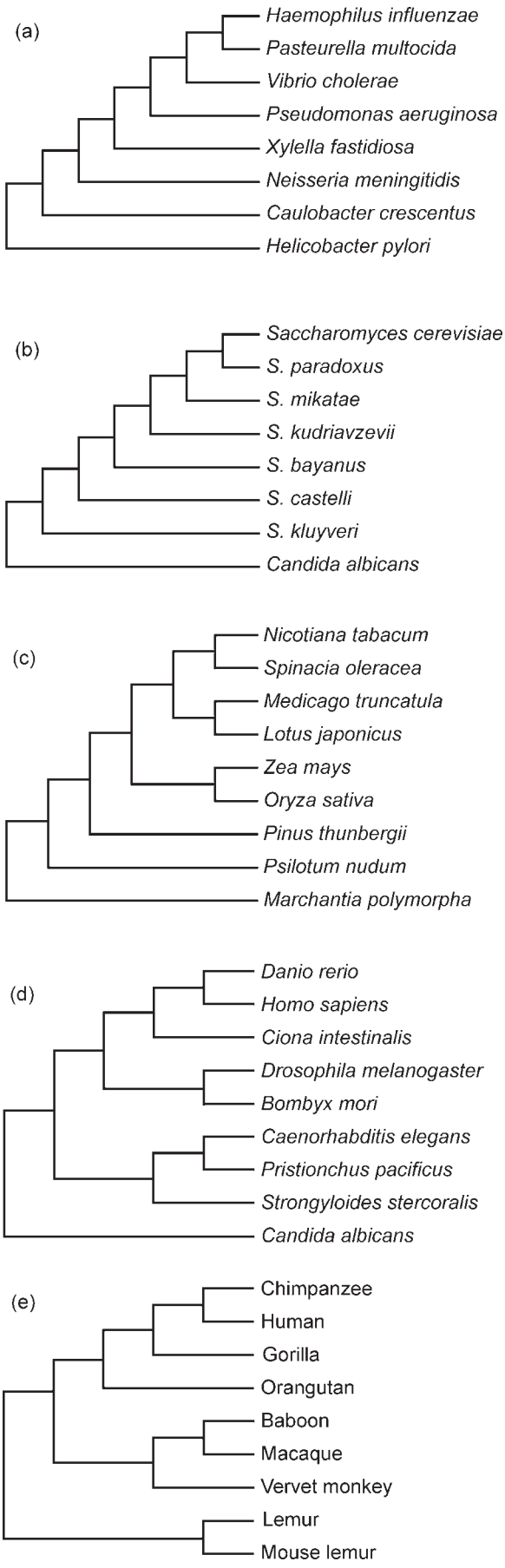


**Figure 4.** The "True" Phylogenies for the Large Datasets

The datasets are as follows: (A) bacterial, (B) yeast, (C) plant, (D) metazoan, and (E) primate.

DOI: 10.1371/journal.pbio.0040088.g004

subset of a larger dataset comprising 730 genes representing 45 species of bacteria [38]. Eight species of Proteobacteria were selected due to their close phylogenetic relationship, as well as representation among the 730 genes. A total of 102 genes spanned all eight of the species that were retained for analysis. The plant dataset was taken from a larger dataset comprising 61 genes from 12 taxa [39]. Nine species were selected in accordance with the stipulation that their phylogeny was known with almost complete certainty. The metazoan alignment was a subset of a larger dataset comprising 123 genes from 36 eukaryotes [40]. Nine metazoan taxa were selected from this dataset so that the tree relating the selected taxa was not in dispute. Genes that were unavailable for one or more of the nine selected taxa were removed, leaving 99 genes in the final dataset used for phylogenetic analysis. The primate dataset was a subset of a 2,160,276 bp alignment of non-coding DNA from 19 mammals [41,42]. The non-primates were removed from the alignment, and sites with a gap in any sequence were removed. The remaining alignment was broken up into 500 alignments of equal length (632 bp). These individual alignments were each intended to represent the data produced by an ordinary phylogenetic study in which a gene fragment has been sequenced from a number of organisms. The question being asked is if we only have one such alignment, how well are we able to reconstruct the phylogenetic relationships of the organisms?

To assess the accuracy of the phylogenetic methods being tested, estimates of the phylogeny need to be tested against the true phylogeny for each dataset. In order to obtain the best possible estimates of the phylogeny for each dataset, the alignments in each of the five datasets were concatenated. The five concatenated alignments were analyzed under the HKY model of nucleotide substitution with gamma-distributed rate variation among sites and a proportion of invariant sites. Each analysis was run for 5,000,000 MCMC steps, with a discarded burn-in of 500,000 steps. Identical trees were obtained using BEAST with a UCLN model and with MrBayes (Figure 4). The trees inferred from the plant, metazoan, and primate datasets agree well with the established trees for these groups. However, the bacterial and yeast phylogenies are relatively uncertain [37,43], and the trees inferred from the concatenated alignments are probably the best estimates currently available. Even if these trees turn out to be different from the true evolutionary histories of the studied organisms, we can at least assume that the trees used in this analysis are very near in tree space to the truth, and therefore we would expect our results to be little affected. The yeast tree inferred in this study from concatenated data agreed with that published by Rokas et al. [37], also confirmed by Phillips et al. [43] under different phylogenetic models.

For each of the five groups of data, each alignment was analyzed using MrBayes (unrooted Felsenstein [UF] model), BEAST with a molecular clock (CLOC); and uncorrelated lognormal relaxed clock (UCLN). The HKY model of nucleotide substitution was assumed, with gamma-distributed

**Table 4.** Accuracy and Precision of Phylogenetic Inference Using Three Bayesian Methods: CLOC, UCLN, and UF

| Dataset | Sample Size | Average Length | Clock Rejected by LRT | Accuracy (%) (True Tree in 95% Credible Set)[a] | | | Precision (Number of Trees in 95% Credible Set)[b] | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | CLOC | UCLN | UF | CLOC | UCLN | UF |
| Bacteria | 102 | 170 aa | 26% | 46.1 | **48.0** | 42.2 | 5.7 | 10.3 | 11.3 |
| Yeast | 106 | 1,198 bp | 76% | 67.0 | **84.9** | 79.2 | 3.5 | 5.9 | 6.5 |
| Plants | 61 | 647 bp | 67% | **91.8** | 88.5 | 83.6 | 7.5 | 15.4 | 9.2 |
| Animals | 99 | 197 aa | 59% | 64.6 | **69.7** | 57.6 | 5.7 | 10.2 | 14.2 |
| Primates | 500 | 632 bp | 13% | 88.8 | **89.0** | 88.8 | 3.1 | 3.4 | 5.1 |

[a]The percentage of alignments for which the 95% credible set contained the ''true'' tree. The numbers in boldface indicate the better performing model.
[b]The largest 10% of credible sets were treated as outliers and excluded from the calculation of the average. The medians were comparatively uninformative and are not given here.
aa, amino acids; LRT, likelihood ratio test.
DOI: 10.1371/journal.pbio.0040088.t004

rate variation among sites and a proportion of invariant sites. Most analyses were run for 500,000 MCMC steps with 50,000 burn-in steps, although some datasets required 1,000,000 steps with 100,000 burn-in steps. All analyses were checked for convergence using the program Tracer 1.2 [23]. The 95% credible set of trees was obtained for each alignment and compared with the "true" trees obtained using the method described above. The accuracy of the methods was considered to be the frequency with which the true tree was contained in the 95% credible set, whereas the average size of the credible sets was taken to represent precision. These terms have statistical definitions, but we take liberties here to facilitate easier interpretation.

All three methods performed poorly in analyses of the bacterial and metazoan datasets. This result is not surprising, however, considering the substantial time depth of these trees. The uncorrelated relaxed-clock method produced the most accurate estimates of phylogeny overall (Table 4). It outperformed other methods in analyses of the bacterial, yeast, metazoan, and primate data, but the molecular clock method was the most accurate in the analysis of the plant data. A large proportion (76%) of the yeast alignments were significantly non-clocklike (as measured by a likelihood ratio test [10] on the true tree topology), which explains the considerable difference in accuracy between the uncorrelated relaxed-clock method (85.8% of the credible sets contain the true tree) and the molecular clock method (67.9% of the credible sets contain the true tree). The superior performance of the molecular clock in the analysis of the plant data, for which the molecular clock was rejected for 67% of the alignments, may be due to the sensitivity of the likelihood ratio test to even small departures from the clock assumption.

In the case of the primate data, all three methods were similarly accurate in estimating phylogenies. This is probably because the data were relatively clocklike, with the molecular clock assumption rejected for less than a third of the alignments. For all of the datasets that were analyzed, the phylogenetic estimates made using a strict molecular clock were the most precise. As expected, the average size of the 95% credible set of trees was always the smallest for the molecular clock method, and nearly always greatest for the unrooted method. Under conditions in which the data more or less conform to a molecular clock, such as the primate data examined in this study, the molecular clock method should be used due to its superior precision.

## Discussion

The relaxed phylogenetics methods described here co-estimate phylogeny and divergence times under a relaxed molecular clock model, thus providing an integrated framework for biologists interested in reconstructing ancestral divergence dates and phylogenetic relationships. The method presented here naturally incorporates the time-dependent nature of the evolutionary process without assuming a strict molecular clock. One of the byproducts of estimating a phylogeny using a relaxed clock is an estimate of the position of the root of the tree, even in the absence of a non-reversible model of substitution [44,45] or a known outgroup.

Recently, a number of authors have begun to investigate the impact of various forms of model misspecification on the accuracy of posterior probabilities of clade support [46–48]. In a Bayesian framework, the absence of a molecular clock assumption (either strict or relaxed) represents a prior belief that the tree topology provides no information about relative branch lengths. We suggest that this represents a poor prior belief, and that Bayesian estimation of phylogeny from short sequences may be biased when the time-dependency of the evolutionary process is not modeled. We would argue that the complex time-dependency of the evolutionary process should not be ignored a priori as has been common practice, but should instead be carefully modeled. This paper represents a first attempt at incorporating a relaxed-clock model into a Bayesian method of phylogenetic inference.

We have presented a large analysis of 102 bacterial, 106 yeast, 61 plant, 99 metazoan, and 500 primate alignments that overall suggests the relaxed-clock models are both more accurate and more precise at estimating phylogenetic relationships than current unrooted methods implemented in MrBayes and other programs. Overall, these initial results suggest that a relaxed phylogenetic approach may be the most appropriate even when phylogenetic relationships are of primary concern and the rooting and dating of the tree are of less interest.

## Materials and Methods

The molecular clock assumption can be relaxed in a variety of ways [13–15,17,49–52]. In Bayesian treatments of the relaxed clock, there is a vector of rates $R=\{r_1,r_2,\ldots,r_{2n-1}\}$ and a corresponding vector of node heights $\mathbf{t}=\{t_1,t_2,\ldots,t_{2n-1}\}$ in units of time. The node height vector, in conjunction with an edge graph, $E$, define an ancestral tree $g=\{E,\mathbf{t}\}$ in units of time. To convert this tree from units of time to molecular evolutionary units, the rates are either assigned to branches [15,17] or to nodes [53,54]. In both types of models, the prior probability of the rates $f_R(R|g)$ can be calculated by the product of the probability of each rate $r_2$ in the tree given the ancestral rate $r_{A(i)}$ and the time $\Delta t_i$ between the ancestral and derived rate:

$$f_R(R|g) = \prod_i f(r_i|r_{A(i)},\Delta t_i). \quad (1)$$

The first such model to be described [15] assigned rates to the midpoints of branches and the assumed lognormal prior distribution relating the midpoint of the ancestral branch to the midpoint of the derived branch. Another interesting model is the exponential distribution model of Aris-Brosou and Yang [17], which employed an exponential prior distribution on rate $r$ with a mean (and therefore standard deviation) equal to the ancestral rate $r_A$, and with no dependence on the time between the two rates. This second model represents a more punctuated view of change in evolutionary rate, so that only the number of branching events, and not the length of time between events, determines the amount of change in evolutionary rate. In all autocorrelated relaxed-clock models, an additional assumption must be made about the rate at the root. For models that assign rates to nodes, it is necessary to treat the root node in a special way, as it does not have a parent node [15]. For models that assign rates to branches, a branch above the root is implied and must be assigned a rate.

In the autocorrelated relaxed-clock models that have been described, including the commonly used lognormal model [15,17,55], it is also necessary to specify the degree of autocorrelation as a prior. Other prior models of rate change, such as the gamma distribution model and the Ornstein-Uhlenbeck process [55], require more than one hyperparameter to be specified, so that selecting suitable values for a particular dataset may be an even more difficult exercise. The effects of varying these hyperparameters are poorly understood [22], but there is likely to be a considerable impact on posterior estimates of rates.

**Uncorrelated relaxed clocks.** We present an alternative to the autocorrelated prior in which there is, a priori, no correlation of the rates on adjacent branches of the tree. Instead we propose a model in which the rate on each branch of the tree is drawn independently and identically from an underlying rate distribution. We investigate two candidates for the rate distribution among branches:

$$r \sim \mathrm{Exp}(\lambda), \quad (2)$$

$$r \sim LogNormal(\mu,\sigma^2). \quad (3)$$

These uncorrelated priors can be framed in a hierarchical Bayesian framework, as with the autocorrelated priors. In this scenario the exponential version of uncorrelated relaxed clock would have a prior probability on the rate vector of:

$$f_R(R|g) = f(R) = \prod_i \lambda e^{-\lambda r_i}. \quad (4)$$

This model corresponds to an exponential prior distribution on rate $r_i$ with a mean (and therefore standard deviation) equal to $\lambda^{-1}$ and no dependence on either the rate of the previous branch or the time between the two rates. The parameter $\lambda$ is a hyperparameter that is fixed and not estimated via MCMC, and represents a prior statement about both the mean and the variance of branch rates. This prior reflects a punctuated view of change in evolutionary rate, so that the prior expectation of the rate at all branches is the same, with no autocorrelation between adjacent branches. Notice that the posterior distribution of rates among branches need not be the same as the prior in this setup and that autocorrelation may exist in the posterior, even though it is not specified in the prior.

Instead of framing Equations 2 and 3 as prior distributions in a hierarchical Bayesian framework, they can instead be reformulated as a full likelihood model. In this case, the branch rates are not independent random variables with a prior distribution, but are instead constrained so as to fit one of the distributions in Equations 2 and 3 exactly. The parameters of the rate distribution are no longer hyperparameters of a prior distribution, but are instead parameters of the likelihood model. This is closely analogous to the common way in which rate heterogeneity among sites is treated [28].

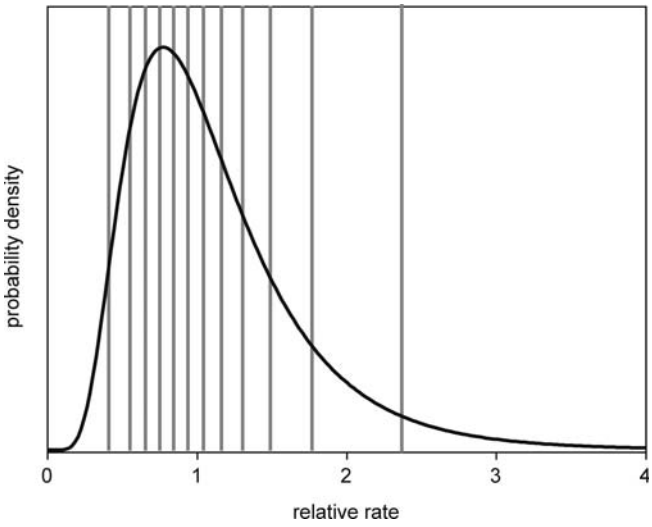


**Figure 5.** A Lognormal Distribution Discretized into 12 Rate Categories
Each of the 12 categories has equal probability ($p = 1/12$). The $i^{th}$ rate category (numbered from left to right) corresponds to the $(I - 0.5)/12$ quantile of the lognormal distribution.
DOI: 10.1371/journal.pbio.0040088.g005

**Priors on phylogeny.** A particular requirement of Bayesian phylogenetic inference is the responsibility given to users to specify a prior probability distribution on the shape of the phylogeny (node ages and branching order). This can be either a benefit or a burden, largely depending on whether an obvious prior distribution presents itself for the data at hand. For example, the coalescent prior [56,57] is a commonly used prior for population-level data and has been extended to include various forms of demographic functions [58,59], sub-divided populations [60], and other complexities. Traditional speciation models such as the Yule process [61] and various birth–death models [62,63] can also provide useful priors for species-level data. Such models generally have a number of hyperparameters (for example, effective population size, growth rate, or speciation and extinction rates), which, under a Bayesian framework, can be sampled to provide a posterior distribution of these potentially interesting biological quantities.

In some cases, the choice of prior on the phylogenetic tree can exert a strong influence on inferences made from a given dataset [64]. The sensitivity of inference results to the prior chosen will be largely dependent on the data analyzed and few general recommendations can be made. It is, however, good practice to perform the MCMC analysis without any data in order to sample wholly from the prior distribution. This distribution can be compared to the posterior distribution for parameters of interest in order to examine the relative influence of the data and the prior (Figure 3).

**Bayesian inference.** The full Bayesian sequence analysis with an uncorrelated relaxed-clock model allows the co-estimation of substitution parameters, relaxed-clock parameters, and the ancestral phylogeny. The posterior distribution is of the following form:

$$f(g,\Theta,\Phi,\Omega|D) = \frac{1}{Z}\Pr\{D|g,\Phi,\Omega\} f_G(g|\Theta) f_{\Theta\Omega\Phi}(\Theta,\Omega,\Phi). \quad (5)$$

The vector $\Phi$ contains the parameters of the relaxed-clock model (e.g., $\mu$ and $\sigma^2$ in the case of lognormally distributed rates among branches). The term $\Pr\{D|g,\Phi,\Omega\}$ is the standard Felsenstein likelihood, where $g$ is a tree with branch length measured in units of time. For the purposes of calculating this likelihood, branch lengths are converted to units of substitutions by multiplying the rates defined by $\Phi$ with the internode distance between node $i$ and parent node $j$ in tree $g$. The tree prior, $f_G(g|\Theta)$, can either be a coalescent-based prior [30,65] for within-population data or some other appropriate prior if the sequences come from multiple populations/species [55]. The vector $\Theta$ contains the hyperparameters of the tree prior. The vector $\Omega$ contains the parameters of the substitution model (such as

transition/transversion ratio, κ; shape parameter for gamma-distributed rates among sites, α; and proportion of invariant sites, $p_{inv}$).

We summarize the posterior density in Equation 5 using samples $(g,\Theta,\Phi,\Omega) \sim f$ obtained via MCMC. If, for example, the divergence times are of primary interest then the other sampled parameters can be thought of as nuisance parameters, and vice versa.

The formulation in Equation 5 implies that the branch-rates could be integrated analytically in the Felsenstein likelihood. Although this could be accomplished relatively easily by discretizing the rate distribution and averaging the likelihood over the rate categories on each branch, we elected to do the integration using MCMC. This was achieved by assigning a unique rate category $c \in \{1,2,\ldots,2n-2\}$ to each branch $j$ of the tree. During the calculation of the likelihood the rate category $c$ is converted to a rate by the following method:

$$r_c = D^{-1}\left(\frac{c-1/2}{2n-2}\right). \quad (6)$$

The function $D^{-1}(x)$ is the inverse function of the probability distribution function, $D(x)=P(X\leq x)$, of the relaxed-clock model specified by Equations 2 and 3. This discretization of the underlying rate distribution is illustrated in Figure 5 for a lognormal distribution with 12 rate categories (sufficient for a tree of seven tips). To integrate the branch rates out, the assignment of rate categories $c$ to branches was sampled via MCMC.

**Model selection.** One issue that remains largely unresolved in this piece of work is the issue of model comparison and model selection. Within a Bayesian framework, Bayes factors are usually regarded as the correct way to deal with model selection. Typically this involves a technique known as reversible-jump MCMC. We have not implemented this, but we do plan on developing a reversible-jump MCMC version of this framework in the future. Typically model selection is easy when one model produces a much better fit. Because all of the models for rate variation examined here differ by one free parameter at most, a simple comparison of the average log posterior probabilities will usually be revealing. It is only when the log posteriors are very similar and the results are qualitatively different between the two models that model selection becomes an issue. This combination of conditions did not occur in any of our real datasets.

**Proposing new states in the MCMC kernel.** The MCMC must sample the tree topology, the divergence times, and the individual parameters of the substitution model and tree prior(s). Therefore, a series of proposal distributions (often called "moves") needs to be employed. Our MCMC implementation employs an array of moves, each of which is designed to explore a certain subspace in the overall parameter/model space being explored. For example, some moves propose local changes to the tree topology while keeping the coalescent interval and all the other parameters constant. Some moves propose a change to a single substitution parameter (such as the shape parameter of the gamma distribution) while keeping everything else constant. The general scheme is to (1) choose a random move with a probability proportional to a specified weight, then (2) apply the move to the current state, and (3) assess the relative score of the new state. The new state is adopted if it has a higher posterior probability; otherwise it is adopted with probability equal to the ratio of its posterior probability to the posterior probability of the previous state. The weights allow the researcher to favor certain moves which can help with the performance of the MCMC, but generally the default weights give good results. Most of the moves used in our MCMC implementation have been previously described [30]. The two new moves involve sampling the rate categories of the branches (a random pair of branches are chosen and their categories are swapped) and dealing with rate categories of branches when a change to the tree topology is made. (We implement two alternatives: keeping all the rate categories the same when a subtree is moved or performing a single rate swap simultaneously with a tree topology change.) These moves are very simplistic, and we suspect that better proposal distributions exist. We have found a small number of datasets in which our current proposal distribution does not work well. Nevertheless, for a large number of datasets including the ones presented in this paper, our scheme performs more than adequately as assessed by repeated runs and estimation of integrated autocorrelation times.

**Summarizing the posterior distribution.** The output of an MCMC analysis is a set of samples from the posterior distribution. In the case of the uncorrelated relaxed-clock models described above, the posterior distribution is a distribution over tree topologies, dates of divergence, branch rates, and parameters of the rate and substitution models. This complex set of samples can be summarized in many ways. One of the simplest summaries of the branch rate distribution is to sample the coefficient of variation ($\sigma_r$; the standard deviation divided by the mean) of the branch rates. Under the exponential model, $\sigma_r = 1$ by definition; under the lognormal model, $\sigma_r$ gives a measure of the degree of clocklikeness of the data. If $\sigma_r = 0$ then the data are perfectly clocklike, whereas larger values correspond to increasing rate heterogeneity among branches. A posterior estimate of $\sigma_r$ can be easily calculated:

$$E[\sigma_r|D] = \frac{1}{L}\sum_{i=1}^{L}\sigma_r^{(i)}. \quad (7)$$

This is the simple average of the calculated $\sigma_r^{(i)}$ over all $L$ samples in the estimated posterior distribution. In addition, 95% HPD limits can also be calculated. In a similar manner, marginal posterior estimates can be calculated for

$E[t_j|D]$ the length of time the $j^{\text{th}}$ branch represents,

$E[r_j|D]$ the rate of evolution on the $j^{\text{th}}$ branch, and,

$E[r_j t_j|D]$ the expected number of substitutions per site occurring on the $j^{\text{th}}$ branch.

Some subtlety in the interpretation of the posterior distribution of rates is required because both the amount of time a branch represents, $t_j$, and the rate of evolution along the branch, $r_j$, are random variables in the MCMC analysis. For example, in general, $E[r_j t_j|D] \neq E[r_j|D]\cdot E[t_j|D]$. For the purposes of this paper, when we refer to the average rate for a set of branches $B$ (such as the set of external branches or the set of internal branches), we define it as the weighted average:

$$r^{(B)} = \sum_{j\in B} r_j t_j \Big/ \sum_{j\in B} t_j, \quad (8)$$

rather than the simple unweighted average $\frac{1}{|B|}\Sigma_{j\in B} r_j$. Thus the posterior estimate of average rate over the whole tree is $E[r^{(T)}|D]$, where $T = \{1,2,\ldots,2n-2\}$. In general, this will be different from the mean of the underlying rate distribution because the rate at each branch is weighted by the time represented by the branch. The justification for this is that the overall rate is best summarized by the total amount of substitutions over the total amount of time, which is what Equation 8 calculates.

**Calibrating the rate of evolution.** In the above discussion on rate models, it was assumed that it is possible to estimate absolute rates of evolution and the variance in absolute rates. In fact, even under a molecular clock assumption, the divergence times and the overall substitution rate can only be separately estimated if there is a source of external calibration information. In the framework described here, this information can come from one of three sources: (1) Prior information on the age of internal nodes: In a phylogenetic context, calibration information is often obtained by assigning the age of a known fossil to a particular internal node [2]. Uncertainty in the association between an internal node and the fossil record can be accommodated by providing a prior probability distribution for the age of the node. Previous studies have used a uniform distribution with upper and lower bounds on the age [54], although other distributions may be suitable [35]. In the above Results section, we presented examples in which calibration times are treated with parametric prior distributions (normal and lognormal). Assigning an age to a particular node is only possible when the tree itself is assumed to be known and fixed, a limitation of previous relaxed-clock implementations [15,17,54]. In the framework presented here, the tree itself is being sampled and thus we cannot define the age of a particular internal node. Instead we specify the age, or the prior distribution of age, for the most recent common ancestor of a set of taxa. Every time a new tree is proposed in the MCMC chain, the most recent common ancestor of the specified taxa is located in the tree, and the prior probability of the age of this node is used to assess the acceptance probability of the proposed tree. (2) Known ages of the sequences: Recently it has also been demonstrated that calibrations can be associated with the sequences at the tips of the tree if they are sampled at significantly different times [29,30,66] with respect to their rate of evolution. Again, there may be uncertainty in calibration dates [67]. The RNA virus data in this study provide examples of this form of calibration information. (3) A strong prior on the substitution rate: If the mean substitution rate is known from a previous study on independent data, then this can be incorporated as prior knowledge. In the simplest case this can be achieved by fixing

the rate of evolution to a known value. It is also straightforward to sample the rate from a parametric distribution obtained from a previous (independent) analysis [68,69]. If there is no prior information about the mean substitution rate, then it can be fixed to 1, resulting in time being in units of substitutions per site.

All of these forms of calibration information can be incorporated into our MCMC implementation either on their own or in any combination, as appropriate.

## Supporting Information

**Protocol S1.** Relaxed Phylogenetics and Dating with Confidence

Found at DOI: 10.1371/journal.pbio.0040088.sd001 (167 KB DOC).

## Acknowledgments

The authors would like to thank S.-M. Chaw and H. Philippe for providing data, and Lindell Bromham for coining the phrase "dating with confidence." All of the methods described above have been implemented in the BEAST software package (http://evolve.zoo.ox.ac.uk/beast).

**Author contributions.** AJD and AR conceived the original idea, developed the software, and performed the marsupial and virus data analyses. SYWH developed the simulation software, performed the simulation analysis, developed the use of prior distributions for calibrating node ages, and performed the analyses on the bacteria, yeast, plant, metazoan, and primate datasets. MJP collected and curated the marsupial dataset and provided expert calibration information. AJD, SYWH, MJP, and AR contributed to the writing of the article.

**Funding.** AJD was supported by the Wellcome Trust. SYWH was supported by a Commonwealth (Oxford) Scholarship from the Commonwealth Scholarship Commission and a Domus Research Studentship and Edward Penley Abraham Cephalosporin Scholarship from Linacre College, Oxford. AR is supported by a University Research Fellowship from The Royal Society.

**Competing interests.** The authors have declared that no competing interests exist. ■

**References**

1. Zuckerkandl E, Pauling L (1962) Molecular disease, evolution and genic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in biochemistry. New York: Academic Press. pp. 189–225.
2. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. pp. 97–166.
3. Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231: 1393–1398.
4. Ayala FJ (1997) Vagaries of the molecular clock. Proc Nat Acad Sci U S A 94: 7776–7783.
5. Hasegawa M, Kishino H (1989) Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jpn J Genet 64: 243–258.
6. Yoder AD, Yang ZH (2000) Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17: 1081–1090.
7. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410.
8. Ho SYW, Jermiin LS (2004) Tracing the decay of the historical signal in biological sequence data. Syst Biol 53: 628–637.
9. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.
10. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368–376.
11. Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) version 3.6 [computer program]. Available: http://evolution.genetics.washington.edu/phylip.html. Accessed 31 January 2006.
12. Swofford DL (2003) PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4 [computer program]. Sunderland (Massachusetts): Sinauer Associates.
13. Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15: 442–448.
14. Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14: 1218–1231.
15. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15: 1647–1657.
16. Hasegawa M, Kishino H, Yano T (1989) Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. J Hum Evol 18: 461–476.
17. Aris-Brosou S, Yang Z (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst Biol 51: 703–714.
18. Cranston K, Rannala B (2005) Closing the gap between rocks and clocks. Heredity 94: 461–462.
19. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109.
20. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21: 1087–1091.
21. Drummond AJ, Rambaut A (2003) BEAST version 1.3 [computer program]. Available: http://evolve.zoo.ox.ac.uk/beast. Accessed 31 January 2006.
22. Ho SY, Phillips MJ, Drummond AJ, Cooper A (2005) Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. Mol Biol Evol 22: 1355–1363.
23. Rambaut A, Drummond AJ (2003) Tracer version 1.2 [computer program]. Available: http://evolve.zoo.ox.ac.uk. Accessed 31 January 2006.
24. Bennett SN, Holmes EC, Chirivella M, Rodriguez DM, Beltran M, et al. (2003) Selection-driven evolution of emergent dengue virus. Mol Biol Evol 20: 1650–1658.
25. Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A 94: 7712–7718.
26. Ferguson NM, Galvani AP, Bush RM (2003) Ecological and immunological determinants of influenza evolution. Nature 422: 428–433.
27. 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.
28. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39: 306–314.
29. Rambaut A (2000) Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16: 395–399.
30. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161: 1307–1320.
31. Amrine-Madsen H, Scally M, Westerman M, Stanhope MJ, Krajewski C, et al. (2003) Nuclear gene sequences provide evidence for the monophyly of australidelphian marsupials. Mol Phylogenet Evol 28: 186–196.
32. Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci U S A 90: 4087–4091.
33. Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. J Mol Evol 43: 610–621.
34. Godthelp H, Wroe S, Archer M (1999) A new marsupial from the Early Eocene Tingamarra local fauna of Murgon, southeastern Queensland: A prototypical Australian marsupial? J Mammal Evol 6: 289–313.
35. Hedges SB, Kumar S (2004) Precision of molecular time estimates. Trends Genet 20: 242–247.
36. Rodriguez F, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142: 485–501.
37. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.
38. Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Res 12: 1080–1090.
39. Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58: 424–441.
40. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, et al. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21: 1740–1752.
41. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101: 13994–14001.
42. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793.
43. Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21: 1455–1458.
44. Yap VB, Speed T (2005) Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol Biol 5: 2.
45. Huelsenbeck JP, Bollback JP, Levine AM (2002) Inferring the root of a phylogenetic tree. Syst Biol 51: 32–43.
46. Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. Syst Biol 53: 265–277.
47. Yang Z, Rannala B (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol 54: 455–470.
48. Buckley TR (2002) Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst Biol 51: 509–523.
49. Cooper A, Penny D (1997) Mass survival of birds across the Cretaceous–Tertiary boundary: Molecular evidence. Science 275: 1109–1113.

50. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol Biol Evol 19: 101–109.
51. Cutler DJ (2000) Estimating divergence times in the presence of an overdispersed molecular clock. Mol Biol Evol 17: 1647–1660.
52. Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12: 823–833.
53. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol 18: 352–361.
54. Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. Syst Biol 51: 689–702.
55. Aris-Brosou S, Yang Z (2003) Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. Mol Biol Evol 20: 1947–1954.
56. Kingman JFC (1982) The coalescent. Stoch Proc Appl 13: 235–248.
57. Kingman JFC (1982) On the genealogy of large populations. J Appl Probab 19A: 27–43.
58. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192.
59. Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155: 1429–1437.
60. Ewing G, Nicholls G, Rodrigo A (2004) Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 168: 2407–2420.
61. Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B 213: 21–87.
62. Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans R Soc Lond B Biol Sci 344: 305–311.
63. Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. Mol Biol Evol 14: 717–724.
64. Welch JJ, Fontanillas E, Bromham L (2005) Molecular dates for the "cambrian explosion": The influence of prior assumptions. Syst Biol 54: 672–678.
65. Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150: 499–510.
66. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. Trends Ecol Evol 18: 481–488.
67. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288: 1789–1796.
68. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, et al. (2004) The molecular population genetics of HIV-1 group O. Genetics 167: 1059–1068.
69. Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A (2003) The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: A Bayesian coalescent approach. Mol Biol Evol 20: 381–387.