ACCEPTED FOR PUBLICATION IN *CURRECT ZOOLOGY*

The Impact of Ancestral Population Size and Incomplete Lineage Sorting on Bayesian Estimation of Species Divergence Times

Konstantinos Angelis and Mario dos Reis*

Department of Genetics, Evolution and Environment, University College London, Gower Street, London, United Kingdom, WC1E 6BT.

Correspondence to:
Mario dos Reis

Department of Genetics, Evolution and Environment

University College London

Darwin Building, Gower Street

London WC1E 6BT, UK

Phone: +44 20 7679 2652

e-mail: mario.barros@ucl.ac.uk

Words: 5,634

Figures: 6

Tables: 4

Abstract

Although the effects of the coalescent process on sequences divergences and genealogies are well understood, the virtual majority of studies that use molecular sequences to estimate times of divergence among species have failed to account for the coalescent process. Here we study the impact of ancestral population size and incomplete lineage sorting on Bayesian estimates of species divergence times under the molecular clock when the inference model ignores the coalescent process. Using a combination of mathematical analysis, computer simulations and analysis of real data, we find that the errors on estimates of times and the molecular rate can be substantial when ancestral populations are large and when there is substantial incomplete lineage sorting. For example, in a simple three-species case, we find that if the most precise fossil calibration is placed on the root of the phylogeny, the age of the internal node is overestimated, while if the most precise calibration is placed on the internal node, then the age of the root is underestimated. In both cases, the molecular rate is overestimated. Using simulations on a phylogeny of nine species, we show that substantial errors in time and rate estimates can be obtained even when dating ancient divergence events. We analyse the hominoid phylogeny and show that estimates of the neutral mutation rate obtained while ignoring the coalescent are too high. Using a coalescent-based technique to obtain geological times of divergence, we obtain estimates of the mutation rate that are within experimental estimates and we also obtain substantially older divergence times within the phylogeny.

Key words: ancestral polymorphism, incomplete lineage sorting, divergence time estimation, gene tree, species tree.

Introduction

The molecular clock hypothesis states that the rate of evolution of molecular sequences is approximately constant with time (Zuckerkandl and Pauling, 1965). This powerful idea means that in practice information from the fossil record can be combined with information from molecular alignments to obtain geological times of divergence for species in a phylogeny. Although the method is widely popular (see Hedges and Kumar, 2009, for an extensive review on applications of the molecular clock to divergence time estimation), most molecular-clock dating studies have ignored the effects of the coalescent process on sequence divergences and thus on divergence time estimates. For example, consider a sample of two nucleotide sequences (genes) belonging to two individuals from a diploid population of N individuals. The expected time to coalescence, that is, the time it takes for the two sequences to find their common ancestor is $2N$ generations (Kingman, 1982a; Kingman, 1982b; Tajima, 1983). On the other hand, if the sequences are sampled from individuals belonging to two different, completely isolated species which diverged T generations ago, then the expected sequence coalescent time is $T + 2N$ (Fig. 1), where N is now the population size of the ancestral species (Gillespie and Langley, 1979). In other words, the divergence time of the genes, T^*, is older than the divergence time of the species (i.e. $T^* > T$), especially so if the size of the ancestral population is large compared to the species divergence time. Furthermore, for sequences sampled from three or more species, the deep coalescence times of the gene sequences mean that sometimes the genealogy of the sequences (the gene tree) will differ from the species tree (Fig. 1), a process known as incomplete lineage sorting (Hudson, 1983; Nichols, 2001). Thus, studies that use the molecular clock to estimate the times of species divergences from molecular data should take into account the effect of ancestral population size and incomplete lineage sorting on gene ages, otherwise biased estimates of species divergence times may be obtained.

Several Bayesian phylogenetic methods have been developed to perform inference under the multi-species coalescent (Rannala and Yang, 2003; Liu and Pearl, 2007; Liu, 2008; Heled and Drummond, 2010; Yang, 2015). However, these methods are computationally expensive and are only practical for small datasets or when using simple nucleotide substitution models. Thus, although the coalescent process has long been recognised as an important aspect of sequence evolution (Takahata et al., 1995; Edwards and Beerli, 2000; Kubatko and Degnan, 2007; Knowles and Kubatko, 2010; Burbrink and Pyron, 2011; Oliver, 2013; Yang, 2014), a majority of molecular clock dating analyses are still carried out ignoring the effects of ancestral population size and incomplete lineage sorting (e.g. Erwin et al., 2011; dos Reis et al., 2012; Jarvis et al., 2014; Misof et al., 2014; Zeng et al., 2014). Furthermore, the biases introduced in time estimates by ignoring the coalescent process do not seem to have been studied.

In this paper we study the impact of ancestral population size and incomplete lineage sorting on Bayesian estimates of divergence times when the coalescent process is ignored. We use a combination of mathematical analysis, computer simulations, and analysis of a real dataset (the hominoid phylogeny) to show that ignoring the coalescent process can have a large impact on estimates of divergence times, even when estimating ancient divergence events. Divergence times can be substantially under or overestimated, depending on the configuration and precision of the fossil calibrations on the tree, with the molecular evolutionary rate being usually overestimated. The problem is severe and this paper highlights an urgent need for the development of efficient, fast computer software that can provide reliable estimates of divergence times under the multi-species coalescent for the large genome-scale datasets now routinely available.

The case of three-species

Here we study the case of estimating the two divergence times in a phylogeny of three species when the coalescent process is ignored. We first provide an approximate mathematical formula for the time and rate estimates and the estimate errors when the amount of molecular data (the number of genes or loci) analyzed is very large, when we have perfect fossil information, and when there is little conflict between the species tree and the sampled gene trees. We then use computer simulations to study Bayesian time estimation when incomplete lineage sorting may be substantial and when we use uncertain fossil calibrations in the form of priors.

A simple approximation to the time and rate estimates and their errors when the coalescent process is ignored

Consider the three-species phylogeny of Figure 1. We wish to estimate the two species divergence times on the tree: $t_1 = gT_1$ (the age of the root) and $t_2 = gT_2$ (the age of the internal node), where T is the time in generations, t the time in years, and g the generation time. Assume that $T_1 – T_2$ is large enough so that the probability of conflict between a gene tree and the species tree is close to zero and can be ignored (the probability of conflict is $P = 2/3 \exp[–(T_1 – T_2)/(2N)]$, so for example, if $T_1 – T_2 > 10N$ then $P < 0.45\%$). In a typical molecular dating analysis, we may sample a set of genes from each species, concatenate and align the sequences (i.e. create a supergene alignment) and then estimate the species phylogeny and the molecular distances using the concatenated alignment. The molecular distances and information from the fossil record are then used to estimate the divergence times. Thus, to understand how time estimates may be affected by ignoring the coalescent process, we must first understand how the molecular distances are affected.

The expected molecular distance (in expected number of substitutions per site) between either of the two genes sampled from A and B and their common ancestor is

$$\mathrm{E}(d_2) = (T_2 + 2N)gr = (t_2 + 2Ng)r, \qquad (1)$$

where r is the substitution rate per site per year (so that $\mu = rg$ is the rate per generation), which we assume to be the same for all loci in all lineages. Similarly, for a sample of two genes from A (or B) and C, the expected distance is

$$\mathrm{E}(d_1) = (T_1+2N)gr = (t_1 + 2Ng)r. \qquad (2)$$

Note that equations (1) and (2) are the expected distances for a pairwise sequence alignment for a single locus. If our supergene alignment is very long (so that it contains a large number of loci) the molecular distances (the branch lengths) estimated on the species tree will be close to the expected values

$$\hat{d}_2 \approx \mathrm{E}(d_2) = (t_2 + 2Ng)r \qquad (3)$$

and

$$\hat{d}_1 \approx \mathrm{E}(d_1) = (t_1 + 2Ng)r\,. \qquad (4)$$

However, as we sample more and more loci, the distance estimates will not converge to the expectations in equations (1) and (2) because they are estimated on the species tree (and not on the pairwise alignments) and incomplete lineage sorting is ignored. We could calculate the correct expectations for $\hat{d}_1$ and $\hat{d}_2$ obtained on the species tree, but this is mathematically tedious and so we work with the approximations of equations (1) and (2) instead (later we will see that the approximations turn out to be quite good even when there is substantial incomplete lineage sorting).

Now imagine that the age t_1 is known (say, from the fossil record). Then under the molecular clock, a naive estimator of t_2 (i.e. naive because it ignores the coalescent process) using t_1 as a calibration is

$$\hat{t}_2 = t_1\hat{d}_2 / \hat{d}_1. \qquad (5)$$

This estimator is constructed under the assumption that the ratio of the species divergence times (t_2/t_1) is the same as the ratio of the molecular distances (d_2/d_1). However, the later ratio

is instead the ratio of *gene* divergence times, and thus the estimator $\hat{t}_2$ will be biased. If we replace the distance estimates in equation (5) with their approximations from equations (3) and (4) we can obtain an approximation for $\hat{t}_2$ as a function of the true parameter values

$$\hat{t}_2 \approx t_1 \frac{(t_2 + 2Ng)}{(t_1 + 2Ng)}. \qquad (6)$$

From equation (6) we can obtain an approximation to the bias of the estimator

$$\hat{t}_2 - t_2 \approx t_1 \frac{(t_2 + 2Ng)}{(t_1 + 2Ng)} - t_2 = \frac{(t_1 - t_2)2Ng}{t_1 + 2Ng}. \qquad (7)$$

The bias in this case is positive and so t_2 is overestimated. On the other hand, if t_2 is known, we could instead construct a naive estimator on t_1 as

$$\hat{t}_1 = t_2 \frac{\hat{d}_1}{\hat{d}_2} \approx t_2 \frac{(t_1 + 2Ng)}{(t_2 + 2Ng)} \qquad (8)$$

which has approximate bias

$$\hat{t}_1 - t_1 \approx \frac{(t_2 - t_1)2Ng}{t_2 + 2Ng}. \quad (9)$$

In this case the bias is always negative and t_1 is underestimated.

We can also construct a naïve estimator on the molecular rate, *r*, for example by using t_1 as the calibration time

$$\hat{r} = \hat{d}_1 / t_1 \qquad (10)$$

Replacing the distance estimate by its approximate expectation we get

$$\hat{r} \approx \frac{(t_1 + 2Ng)r}{t_1} = r + \frac{2Ngr}{t_1}. \qquad (11)$$

Thus the approximate bias of the rate estimator is

$$\hat{r} - r \approx 2Ngr / t_1. \qquad (12)$$

The bias is always positive and so *r* is overestimated. Similarly we can obtain a rate estimate using t_2 as the calibration time

$$\hat{r}' = \hat{d}_2 / t_2 \approx r + 2Ngr / t_2, \qquad (13)$$

with approximate bias

$$\hat{r}' - r \approx 2Ngr / t_2. \qquad (14)$$

The bias here is also positive and so the rate is overestimated. However, in this case the overestimation is more severe than when using t_1 as the calibration time.

The relative error of an estimator is the bias of the estimator divided by the true parameter value

$$\varepsilon(\hat{t}) = (\hat{t} - t) / t. \qquad (15)$$

We can use the biases of equations (7), (9) and (12) to obtain approximations to the relative errors on the estimates of t_1, t_2 and r. Note that if the relative error is positive, then the parameter is overestimated, and if it is negative the parameter is underestimated. Figures 2 and 3 show the relative errors on estimates of t_1, t_2 and r for a few cases when the coalescent process is ignored. The errors can be substantial. For example, when t_2 = 1 million years ago (Ma), t_1 = 10 Ma, g = 10 years (y) and $N = 10^5$ individuals, t_2 is overestimated by 150% when using t_1 as the calibration time (Figure 2A). On the other hand, for the same parameter values and when t_2 is used as the calibration time, t_1 is underestimated by 60% (Figure 2B) and r is overestimated by 200% (Figure 3).

Simulation analysis: Bayesian estimates of times when the coalescent process is ignored

We simulated samples of 50 loci (each locus of length 1,000 nucleotides) from a three-species phylogeny (Figure 1) using the program MCCOAL (Rannala and Yang, 2003; Yang and Rannala, 2010). MCCOAL simulates gene trees under the multi-species coalescent with corresponding gene alignments (the alignments are generated under the Jukes and Cantor, 1969, substitution model, JC69). The species divergence times are t_1 = 10 Ma, and t_2 = 1, 5, 9 Ma. The generation time is g = 10 years, and the substitution rate is $r = 10^{-9}$ substitutions per

site per year (s/s/y). The population size (assumed to be constant in all lineages) is $N = 10^2$, 10^3, 10^4, 10^5, 10^6 individuals. This gives a total of 3×5 = 15 parameter combinations. The number of replicates (the number of times each parameter setup is simulated) is 100.

The simulated alignments were concatenated into a supergene alignment, and Bayesian estimates of divergence times under the clock and the JC69 model where obtained using the program MCMCTREE (Yang, 2007). The species tree is used and assumed known. We set one time unit to be 10 My. The substitution rate has a gamma prior, $r \sim G(1, 100)$, with mean 0.01 per time unit (i.e, meaning 10^{-9} s/s/y). We used two strategies to construct the time prior. Strategy 1: The prior on the age of the root is $t_1 \sim G(100, 100)$. This is an informative prior, equivalent to a fossil calibration with mean 10 Ma and 95% prior interval 8–12 Ma. For t_2 we use a diffuse prior density conditioned on t_1 (a uniform distribution between 0 and t_1). Strategy 2: We use informative calibrations on both times, $t_1 \sim B(0.7, 1.4)$ and $t_2 \sim B(0.4, 0.6)$, equivalent to 7–14 My and 4–6 Ma respectively. Here B(*a*, *b*) means that the time is uniformly distributed beetween a minimu age *a* and a maximum age *b*, but with 5% probability that the time is outside the interval (i.e., 2.5% on each side). Note the calibration on t_2 has less uncertainty (with the uncertainty measured by the calibration width divided by the midpoint of the calibration, as in dos Reis and Yang, 2013) than that on t_1. The simulated data, *D*, were analysed under both calibration strategies, and the posterior mean of the times and rate, $\tilde{t} = \mathrm{E}(t \mid D)$ and $\tilde{r} = \mathrm{E}(r \mid D)$, their relative errors, and the 95% credibility intervals (CIs) were collected and averaged among the 100 replicates.

Table 1 shows a few summary statistics for the simulated data sets. The amount of incomplete lineage sorting (i.e. the probability of conflict between gene trees and the species tree) varied from 0% (for t_2 = 1 Ma, t_1 = 10 Ma and $N = 10^2$) up to 63.4% (for t_2 = 9 Ma, t_1 = 10 Ma and $N = 10^6$). Table 1 also shows the maximum likelihood estimates (MLE) of the molecular distances for the supergene alignment obtained on the species tree using the

program BASEML (Yang, 2007). The estimated distances are virtually identical to the expectations (equations 1 and 2) when incomplete lineage sorting is negligible; and they are still very close to the expectations even when incomplete lineage sorting is substantial (bold lines in Table 1). This shows that the long supergene alignment is very informative about the molecular distances (i.e. there is little error in the MLE of the distances).

Figure 4 and Table 2 show Bayesian estimates of times and the molecular rate as a function of the population size for the simulated data. Under calibration strategy 1, the age of the root is correctly recovered in all cases owing to the informative calibration on t_1. On the other hand, t_2 is overestimated, with the estimate's error becoming increasingly worse with increasing population size (Figure 4A-4C). For example, for $N = 10^6$ and $t_2 = 1$ Ma, $\tilde{t}_2$ is 7.28 Ma, i.e. a relative error of 628% (Figure 4A). The rate is also overestimated as N increases, irrespective of the true age of the internal node. For example, for $N = 10^6$, $\tilde{r} = 2.9\times10^{-9}$ s/s/y (relative error 190%) for $t_2 = 5$ Ma (Figure 4B', Table 2). In calibration strategy 2, t_2 has the most precise (or informative) calibration, and so this calibration dominates the analysis. The age of the root in this case is underestimated, and the rate overestimated, as N increases.

The posterior time and rate estimates in Figure 4 are close to the approximations for the naïve estimators (solid line) calculated with equations (6), (8), (11) and (13). Note that the naïve estimates of times and rate were derived without reference to any particular nucleotide substitution model. Thus the theory of equations (6), (8), (11) and (13) is also expected to apply to simulations carried out under more complex substitution models such as HKY or GTR (Hasegawa et al., 1985; Yang, 1994), that is, we expect to see the same biases and relative errors on the estimates. The use of JC69 in our simulations here is thus unimportant, and has no bearing on the properties of sequence evolution under the multi-species coalescent.

Simulation analysis: Bayesian estimates of times under the multi-species coalescent

We re-analyzed the simulated gene alignments on the three-species phylogeny with the program BPP (Yang, 2015), which can be used to obtain estimates of relative divergence times among species, τ, under the multi-species coalescent. The relative times are given as expected number of substitutions per site (i.e. they are the molecular distances between the tips of the phylogeny and the species divergence events, so that $\tau = rt$), and so we devise a method to translate these relative times into geological time estimates. In this section we aim to highlight how analysis under the correct model (the multi-species coalescent) can produce time estimates that are unbiased and have little error.

Note that with BPP, the gene alignments *are not* concatenated. Sequences are analysed under the JC69 model and under the clock. We assigned a gamma prior on the relative age of the root, $\tau_1 \sim G(2, 200)$, with mean 0.01 (the true value of τ_1). For τ_2 we used a diffuse prior conditioned on τ_1 (uniform between 0 and τ_1). For the population size parameters, $\theta = 4N\mu$, we used a gamma prior, $\theta \sim G(2, \beta)$, with β set so that the mean of the distribution matches the true population size in the simulations. BPP estimates one θ per ancestral lineage (i.e. two values for the phylogeny of Figure 1, one for the AB ancestral lineage, and another for the ABC lineage beyond the root). We assumed the same mutation rate across loci.

The relative divergence times estimated with BPP can be translated into absolute geological times by using either a fossil calibration or a prior on the per year mutation rate, r. We used the following procedure. Consider an MCMC sample from the posterior distribution of relative ages (i.e., the i-th sample of the relative root age is τ_1^i) obtained with BPP. First, we sampled values, t_1^i, from a prior density on the root age $t_1 \sim G(100, 100)$. Then samples for the age of the internal node and the per year mutation rate are given by $t_2^i = t_1^i \tau_2^i / \tau_1^i$ and $r_i = \tau_1^i / t_1^i$, respectively. We simply sampled as many values of t_1^i as the number of samples in the MCMC. In this way we obtain a posterior sample of t_1, t_2 and r under the multi-species

coalescent (the posterior of t_1 is simply the prior sampling density). The resulting sample can be summarised in the usual way to obtain the posterior mean of times, rate and 95% CIs.

Divergence times estimated with BPP for the case t_2 = 5 Ma are shown on Table 2. The posterior means for t_2 and r are very accurate (close to the true values) with little relative error. Furthermore, the 95% CIs always contain the true values, even when the mismatch probability between gene trees and the species tree is high. However, for large population sizes the uncertainty around $\tilde{t}_2$ can be quite large because of substantial variation in the coalescent times across genes. For example, for $N = 10^6$ the CI is 2.43−8.12 Ma. Estimates for the cases t_2 = 1 and t_2 = 9 Ma show similar trends (high accuracy and low error) and are not shown.

The case of nine-species

In the three-species case we saw that the molecular rate is overestimated when the coalescent process is ignored and that time estimates may be under or over estimated depending on which node has the most precise fossil calibration. For phylogenies of more than three-species with multiple fossil calibrations the situation is expected to be more complicated. We use computer simulation and Bayesian analysis to study time estimation on a nine-species phylogeny with multiple fossil calibrations.

We simulated gene samples for 50 loci (each 1,000 nucleotides long) on the nine-species phylogeny of Figure 5 using the MCCOAL program. We considered two cases: (1) a young phylogeny where the root is 10 Ma; and (2) an old phylogeny where the root is 100 Ma. The true ages of the internal nodes are shown in Figure 5. In both cases the true substitution rate is $r = 10^{-9}$ s/s/y, and the generation time is g = 10 y. The true times and the mutation rate are used to calculate the relative ages, τ, needed by the program MCCOAL in the simulation. For example, in the young phylogeny the relative age of the root is τ_{10} = 10 My × 0.001 s/s/My = 0.01 s/s, while in the old phylogeny it is τ_{10} = 100 My × 0.001 s/s/My = 0.1 s/s. We simulated

gene alignments on the two phylogenies assuming a constant population size in all lineages, with $N = 10^3$, 10^4, 10^5, 10^6 individuals. Additionally, we simulated a more realistic case where N varied among lineages (between 10^3 and 10^6 individuals, Figure 5). In total we simulated 10 cases (2 phylogenies × 5 population size cases). The number of simulation replicates was 100.

The simulated alignments were concatenated into a supergene alignment and analyzed with the MCMCTREE program to estimate the species divergence times and the rate. Analyses were carried out under the clock and under the JC69 model of nucleotide substitution. The parameters of the birth-death model with species sampling (used to specify the time prior on nodes without fossil calibrations) were set to $\lambda = \mu = 1$ and $\rho = 0$ (Yang and Rannala, 2006). These values specify a diffuse uniform kernel density on the node ages. One time unit was set to be 10 My for the young phylogeny and 100My for the old one. We used diffuse priors on the rate: $r \sim G(1, 100)$ and $r \sim G(1, 10)$ for the young and old phylogenies respectively, with prior means 0.01 and 0.1 substitutions per time unit, with both meaning 10^{-9} s/s/y. Four nodes have soft fossil calibrations: $t_{10} \sim$ B(0.5, 1.5), $t_{12} \sim$ B(0.1, 0.3), $t_{13} \sim$ B(0.3, 0.5) and $t_{16} \sim$ B(0.2, 0.4), where, for example, B(0.5, 1.5) means that the divergence time is between 5 and 15 Ma in the young phylogeny, or between 50 and 150 Ma in the old phylogeny (Yang and Rannala, 2006). The posterior mean of the times and rate, their relative errors, and the 95% CIs were collected and averaged among the 100 replicates.

Table 3 shows the posterior means of times and rate averaged across all replicates. For the young phylogeny, when N is small (10^3 and 10^4) there is no incomplete lineage sorting ($P = 0$ %) and node ages are overestimated, with relative errors ranging from 5% to 20% (Table 3). As N increases (from 10^5 to 10^6) the ages of nodes close to the root (t_{10}, t_{11}, t_{13}, t_{15}) become increasingly underestimated, while the ages of the external nodes (t_{12}, t_{14}, t_{16}, t_{17}) become increasingly overestimated. For the larger N values the amount of incomplete lineage sorting is substantial and the relative errors in time and rate estimates can be quite dramatic. For

example, for $N = 10^6$, the age of the root is underestimated from 10 Ma to 4.8 Ma (–52% error) while the age of a young node (17) is overestimated from 1 Ma to 2.6 Ma (160%), and the molecular rate is overestimated by 628% (Table 3). Similar trends can be noticed when N varies among lineages. For example, the ages for external nodes with large ancestral population sizes (i.e. nodes 12 and 17) were overestimated, while the ages for external nodes with small ancestral population sizes (i.e. nodes 14, 16) where underestimated, as did the ages for the nodes close to the root (i.e. nodes 10, 11, 15) (Table 3).

Results for the old phylogeny were similar to the young phylogeny, although the errors in the estimates are smaller (Table 3). This is because in the old phylogeny there is substantially less incomplete lineage sorting, and the discrepancies between gene divergence times and species divergence times are less severe. For example, for $N = 10^6$ and $g = 10$ y we expect genes to coalesce at $2Ng = 20$ My over the speciation event, so genes that enter the ancestral population at the root of the phylogeny, would have an expected divergence time of 120 Ma in the old phylogeny, or 20% older than the root speciation event at 100 My. However, for the small phylogeny, the equivalent case means an expected gene divergence age of 30 Ma, or 200% older than the root speciation event at 10 My. This conflict between gene ages and species ages clearly leads to the errors in the divergence time estimates. Note that although the situation is not as severe in the old phylogeny, the relative errors are still substantial. For example, in the old phylogeny, for $N = 10^6$, the relative error on the age of the root is –20.5%, while for one of the younger nodes (node 14) the error is 93%, and for the molecular rate the error is 54% (Table 3). For both the young and old phylogenies, the molecular rate is overestimated when the amount of incomplete lineage sorting is substantial (Table 3).

Divergence times of four hominoid species

We now study the discrepancies in time and rate estimates when they are estimated under the multi-species coalescent vs. estimates obtained when ignoring the coalescent in a real data set.

We use the hominoid phylogeny of Figure 6 as a case study. The molecular data are from Burgess and Yang (2008) and consist of 14,663 neutrally evolving loci. First we estimated the divergence times and rate ignoring the coalescent process, that is, by using the program MCMCTREE. Then we re-analysed the data under the multi-species coalescent, that is, by using the program BPP.

For the MCMCTREE analysis all loci were concatenated into a single alignment and the analysis was carried out under the JC69 substitution model and the strict clock. The time unit was set to 10 My. We used a diffuse gamma prior $r \sim G(1, 100)$ with mean 0.01 (meaning 10^{-9} s/s/y). The fossil calibrations are from dos Reis et al. (2012) and are shown in Figure 6. We used an upper bound of 33.7 Ma for the human-gorilla split while only a minimum bound had been used by dos Reis et al. (2012).

In the BPP analysis the multi-locus sequence data were analyzed assuming the same mutation rates across loci. A gamma prior was used for the population size parameters, $\theta \sim G(2, 500)$, with mean 0.004. The relative age of the root in the species tree (τ_{HCGO}) was assigned a gamma prior $G(4, 219)$, with mean 0.018, while the other relative divergence times were assigned a diffuse Dirichlet prior conditioned on τ_{HCGO}. The prior mean for τ_{HCGO} was set based on a divergence time for the human-orangutan split of 18.3 Ma (Steiper and Young, 2006) and a mutation rate equal to 10^{-9} s/s/y. The relative divergence times obtained with BPP were translated into geological times. We used two calibrations: (1) Values for the age of the human-chimp divergence, t_{HC}, were sampled from a uniform distribution between 5.7 and 10 Ma (equal to the fossil calibration for this node used with MCMCTREE). Then the sampled values were used to calculate samples for the ages of the other nodes and the molecular rate (e.g. $t_{HCG} = t_{HC} \times \tau_{HCG}/\tau_{HC}$ and $r = \tau_{HC}/t_{HC}$). (2) Alternatively, values for the molecular rate, r, were sampled from a gamma distribution $G(100, 200)$ with mean 0.5 and 95% prior interval 0.4-0.6 (meaning 0.4 to 0.6×10^{-9} s/s/y). This distribution is based on experimental estimates

of *de novo* mutation rates in the human genome (see Scally and Durbin 2012). The sampled rate values were then used to calculate the divergence times (i.e. $t_{HC} = \tau_{HC}/r$). Similarly, we obtained estimates of ancestral population size (N) by sampling from θ and assuming a generation time of 20 years for the ancestral hominoid lineages (Langergraber et al., 2012; Scally and Durbin, 2012).

Bayesian estimates of divergence times and the molecular rate obtained with the MCMCTREE and BPP programs are shown in Table 4. The posterior mean of the molecular rate obtained with MCMCTREE, $\tilde{r} = 0.8 \times 10^{-9}$ s/s/y, is higher than the estimate obtained with BPP, $\tilde{r} = 0.53 \times 10^{-9}$ s/s/y, when the human-chimp split is used to calibrate the phylogeny (Table 4). The BPP estimate is well within the 0.4×10^{-9} to 0.6×10^{-9} s/s/y range from *de novo* mutation experiments (Scally and Durbin, 2012).

Note that the uncertainties (or relative errors) of the MCMCTREE calibrations are 54.8%, 129% and 100% for the human-chimp, the human-gorilla, and the human-orangutan calibrations respectively, where the uncertainty is measured as the (calibration width)/(calibration midpoint) (dos Reis and Yang, 2013). Thus, the human-chimp calibration is by far the most precise, and it is thus the most informative about estimation of the molecular rate. Given that there is substantial ancestral polymorphism and incomplete lineage sorting in the ape phylogeny (Burgess and Yang, 2008), the estimated molecular rate by MCMCTREE is almost surely an overestimate. We can use the naive estimate of eq. (13) to gain insight into the overestimation error on the molecular rate. For example, assuming a true mutation rate of 0.5×10^{-9} s/s/y (Scally and Durbin, 2012), a true divergence time for human-chimp of 7.85 Ma (the midpoint of the calibration), an ancestral population size of 150,000 individuals (Table 4), and a generation time of 20 y (Langergraber et al., 2012), we get that the naive rate estimate (eq. 13) ignoring the coalescent process is $\hat{r}' \approx 0.88 \times 10^{-9}$ s/s/y, which is reasonably close to the posterior mean of $\tilde{r} = 0.8 \times 10^{-9}$ s/s/y obtained with MCMCTREE.

In other words, the expected discrepancies between gene ages and species ages under the coalescent can be used to explain the observed discrepancies in rate estimates between BPP and MCMCTREE.

Time estimates for the human-gorilla and human-orangutan divergences obtained with BPP are substantially older than those obtained with MCMCTREE. For example, the posterior means of the human-orangutan divergence obtained with BPP at 26.6 Ma and 27.8 Ma are 16% and 21% older than that obtained with MCMCTREE at 22.9 My (Table 4). Thus the MCMCTREE estimate of the root age is likely an underestimate. Note that time estimates obtained with BPP under the rate calibration are the most precise (i.e. they have the narrower 95% CI width). This is because the rate calibration has less uncertainty than the human-chimp calibration: 40% vs. 54.8% respectively. Assuming that the rate calibration is correct (i.e. that the *de novo* mutation rate measurements are accurate) then the time estimates under the BPP rate calibration would be the most accurate and should be preferred.

Discussion

Results from our theoretical, simulation, and real data analyses indicate that polymorphism in ancestral lineages and incomplete lineage sorting can significantly affect Bayesian estimates of divergence times and of the molecular evolutionary rate when the inference models do not account for the multi-species coalescent. Whether times are over or underestimated depends on the relative precision and configuration of the fossil calibrations on the tree. If very precise calibrations are used on young nodes on a phylogeny, the ages of ancient divergence times can be grossly underestimated. Note that this is expected to occur even in ancient phylogenies. For example, if the age of a young node and the ancestral population size are such that the gene divergence time is twice the age of the young node, then the molecular evolutionary rate will be overestimated by 100% (i.e. it will be roughly twice the true value), and the age of the root of the phylogeny will be underestimated by 50% (i.e. it will be half the

true value), independently of how ancient the root divergence event is. On the other hand, if the most precise calibrations are placed on the most ancient nodes of a phylogeny (perhaps a less common case), then the ages of the younger nodes in the phylogeny will tend to be overestimated. In both cases the molecular rate will tend to be overestimated.

Note that in our Bayesian analyses with the MCMCTREE program the sequence data were analysed as a single concatenated alignment. Alternatively we could separate each locus into individual partitions (or group into several partitions) and estimate the divergence times assuming variable rates among loci. This approach is not expected to affect time estimates and their errors because inference is done under the strict molecular clock and the species phylogeny is assumed known and the same for all loci. Indeed we re-analysed the simulated data in the nine-species phylogeny with MCMCTREE by allowing each locus to evolve according to its own substitution rate. The time estimates were virtually identical to those for the concatenated alignment. On the other hand, in the new analyses we obtained individual rate estimates for each locus, with the rate estimates being overestimated and following a distribution centred around the single rate estimate for the concatenated analysis.

Here we assumed that species were completely separated after speciation, with no gene flow between the novel species after the speciation event. This is clearly an unrealistic assumption and the effect of this on divergence time estimates requires further work. An additional assumption of the multi-species coalescent is that the sequences sampled are neutrally evolving (like the set of sequences analysed for the hominoid phylogeny, Burgess and Yang, 2008). Episodes of positive selection may affect the relative ages of gene coalescent events and may affect divergence time estimates. More work will be required to address this issue.

Although the effect of the coalescent process on sequence evolution is well understood, the vast majority of molecular clock dating studies have ignored the issue (e.g. dos Reis et al.,

2012; Jarvis et al., 2014), perhaps under the innocent belief that incomplete lineage sorting and ancestral polymorphism should only be taken into account when analysing closely related species. Our results here highlight that the problem is much worse and that the coalescent process should be incorporated into analyses of divergence times at all timescales. Unfortunately, software currently available to perform Bayesian phylogenetic inference under the multi-species coalescent is either computationally expensive (e.g. *BEAST, Heled and Drummond, 2010) or has been designed to work only for closely related sequences (e.g. BPP, Yang, 2015). For example, we chose to analyze the hominoid phylogeny because the BPP program can only perform inference under the strict molecular clock and under the JC69 substitution model. These assumptions are met in the hominoid phylogeny: The clock is not violated and the molecular distances are small enough so that the JC69 model can adequately describe the substitution process. In order to analyze more ancient phylogenies, multi-species coalescent models that incorporate molecular rate variation among lineages, that use more complex substitution models, and that can handle the large amounts of genomic data now available will be required.

Acknowledgements

We thank Ziheng Yang and two anonymous reviewers for valuable comments. KA was supported by a University College London Impact Studentship (UK). MdR was supported by Biotechnology and Biological Sciences Research Council (UK) grant BB/J009709/1 awarded to ZY.

References

Burbrink FT, Pyron RA, 2011. The impact of gene-tree/species-tree discordance on diversification-rate estimation. Evolution. 65: 1851-1861.

Burgess R, Yang Z, 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 25: 1979-1994.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, et al., 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci. 279: 3491-3500.

dos Reis M, Yang Z, 2013. The unbearable uncertainty of Bayesian divergence time estimation. J Syst Evol. 51: 30-43.

Edwards SV, Beerli P, 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution. 54: 1839-1854.

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, et al., 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science. 334: 1091-1097.

Gillespie JH, Langley CH, 1979. Are evolutionary rates really variable? J. Mol. Evol.: 27-34.

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.

Hedges SB, Kumar S, 2009. The timetree of life. Oxford University Press, Oxford.

Heled J, Drummond AJ, 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27: 570-580.

Hudson RR, 1983. Properties of a neutral allele model with intragenic recombination. Theoet. Popul. Biol.: 183-201.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346: 1320-1331.

Jukes TH, Cantor CR, 1969. Evolution of protein molecules. Mammalian Protein Metabolism, edited by H. N. Munro. Academic Press, New York: 21-123.

Kingman JF, 1982a. The coalescent. Stochastic Process Appl.: 235–248.

Kingman JF, 1982b. On the genealogy of large populations. J. Appl. Prob.: 27-43.

Knowles LL, Kubatko LS, 2010. Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, Hoboken, N.J.

Kubatko LS, Degnan JH, 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 56: 17-24.

Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, et al., 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci U S A. 109: 15716-15721.

Liu L, 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24: 2542-2543.

Liu L, Pearl DK, 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56: 504-514.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, et al., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346: 763-767.

Nichols R, 2001. Gene trees and species trees are not the same. Trends Ecol Evol. 16: 358-364.

Oliver JC, 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution. 67: 1823-1830.

Rannala B, Yang Z, 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164: 1645-1656.

Scally A, Durbin R, 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13: 745-753.

Steiper ME, Young NM, 2006. Primate molecular divergence dates. Mol Phylogenet Evol. 41: 384-394.

Tajima F, 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105: 437-460.

Takahata N, Satta Y, Klein J, 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol. 48: 198-221.

Yang Z, 1994. Estimating the pattern of nucleotide substitution. J Mol Evol. 39: 105-111.

Yang Z, 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586-1591.

Yang Z, 2014. Molecular evolution: a statistical approach. Oxford University Press, Oxford.

Yang Z, 2015. BPP 3.1: A Bayesian MCMC program for analysis of genomic sequence data under the multispecies coalescent model. Curr. Zool.: this issue.

Yang Z, Rannala B, 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol. 23: 212-226.

Yang Z, Rannala B, 2010. Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci U S A. 107: 9264-9269.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, et al., 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun. 5: 4956.

Zuckerkandl E, Pauling L, 1965. Evolutionary divergence and convergence in proteins. In: Bryson V and Vogel HJ (eds). Evolving Genes and Proteins. New York: Academic Press, 97-166.

Figures

Figure 1. A three-species phylogeny. The species tree is represented by thick black lines. The grey lines represent the genealogy for a sample of three genes (one from each species) that matches the species tree. The green dashed lines represent a gene genealogy that does not match the species tree (i.e. we say the species tree and the gene tree are in conflict). If the species have been completely isolated since divergence (i.e. no migration or introgression), then the gene divergence times (T*) will always be older than the species divergence times (T). The expected gene divergence time (in generations) is $E(T^*) = T + 2N$, where N is the size of the ancestral population.

Figure 2. Relative errors in estimates of divergence times on a three-species phylogeny as a function of population size when the coalescent process is ignored. The errors are calculated approximately using equations (7), (9) and (15). (A) Relative errors of estimates of the internal node's age, t_2, when the age of the root, t_1, is known and used as the calibration. (B) Relative errors of t_1 estimates when t_2 is the calibration. In (A) and (B) the true values are t_1 = 10 Ma, t_2 = 1 or = 5 Ma, and g = 10 y.

Figure 3. Relative errors in estimates of the molecular substitution rate on a three-species phylogeny as a function of population size (Figure 1) when the coalescent process is ignored. The error is calculated approximately using equations (14) and (15), with t_2 known and used as the calibration. The true values are t_2 = 1 or = 5 Ma, g = 10 y, and $r = 10^{-9}$ s/s/y.

Figure 4. Bayesian estimates of divergence times (A-D) and the molecular rate (A'-D') for simulated data on a three-species phylogeny. The data were simulated under the multi-species coalescent, but the coalescent process is ignored during Bayesian estimation of divergence

times with the program MCMCTREE. In all cases the true rate is $r = 0.001$ s/s/My. In (A-C) and (A'-C') the root has the most precise calibration, $t_1 \sim G(100, 100)$ while the internal node has a diffuse prior density, $t_2|t_1 \sim U(0, t_1)$. In these cases the age of the root is correctly estimated, but the age of the internal node and the molecular rate are both progressively overestimated with larger N. In (D, D') the internal node has the most precise calibration, $t_2 \sim B(0.4, 0.6)$ vs. $t_1 \sim B(0.7, 1.4)$. In this case the age of the root is progressively underestimated, and the molecular rate is overestimated, with larger N. The solid lines indicate estimates for t_2 (in A-C) or t_1 (in D) and r (in A'-D') calculated using the naïve estimators of equations (6), (8) and (11, 13), respectively.

Figure 5. A nine-species phylogeny used to simulate gene alignments under the multi-species coalescent. The fossil constraints used for Bayesian estimation of divergence times are shown as dotted bars. The numbers in brackets correspond to the ancestral population sizes, N, for the corresponding branches for the case of variable N among lineages.

Figure 6. The phylogeny of four hominoid species showing the fossil calibrations used for time estimation with the program MCMCTREE. The fossil calibrations are soft, i.e., there is a 2.5% probability that the divergence time lies outside the bounds.

Tables

Table 1. Naïve estimates of divergence times and their errors as a function of population size in a three-species phylogeny.

t_2	N	P	E(d_2)	$\hat{d}_2$	E(d_1)	$\hat{d}_1$	$\hat{t}_2$	$\varepsilon(\hat{t}_2)$	$\hat{t}_1$	$\varepsilon(\hat{t}_1)$
1	10^2	0.000	0.0010	0.0010	0.0100	0.0100	1.00	0%	9.98	–0.2%
	10^3	0.000	0.0010	0.0010	0.0100	0.0100	1.02	2%	9.82	–1.8%
	10^4	0.000	0.0012	0.0012	0.0102	0.0102	1.18	18%	8.50	–15.0%
	10^5	0.007	0.0030	0.0030	0.0120	0.0120	2.50	150%	4.00	–60.0%
	10^6	**0.425**	**0.0210**	**0.0211**	**0.0300**	**0.0297**	**7.00**	**600%**	**1.43**	**–85.7%**
5	10^2	0.000	0.0050	0.0050	0.0100	0.0100	5.00	0%	10.00	0%
	10^3	0.000	0.0050	0.0050	0.0100	0.0100	5.01	0.2%	9.98	–0.2%
	10^4	0.000	0.0052	0.0052	0.0102	0.0102	5.10	2.0%	9.81	–1.9%
	10^5	0.055	0.0070	0.0070	0.0120	0.0120	5.83	16.6%	8.57	–14.3%
	10^6	**0.519**	**0.0250**	**0.0243**	**0.0300**	**0.0294**	**8.33**	**66.6%**	**6.00**	**–40%**
9	10^2	0.000	0.0090	0.0090	0.0100	0.0100	9.00	0%	10.00	0%
	10^3	0.000	0.0090	0.0090	0.0100	0.0100	9.00	0%	10.00	0%
	10^4	0.005	0.0092	0.0092	0.0102	0.0102	9.02	0.2%	9.98	–0.2%
	10^5	0.404	0.0110	0.0110	0.0120	0.0120	9.17	1.9%	9.82	–1.8%
	10^6	**0.634**	**0.0290**	**0.0280**	**0.0300**	**0.0295**	**9.67**	**7.4%**	**9.31**	**–6.9%**

Note.− Times are in My. The age of the root is t_1 = 10 Ma, and the generation time is g = 10 y. The time estimates are calculated using equations (6) and (8), and the relative errors with equations (7) and (9). $P = 2/3 \exp[-(T_1 - T_2)/(2N)]$ is the species tree-gene tree mismatch probability. E(d_2) and E(d_1) are the expected molecular distances from the tips of the phylogeny to the respective coalescent events (equations 1 and 2). The molecular distance estimates, $\hat{d}_2$ and $\hat{d}_1$, are obtained from data simulated with the program MCCOAL and estimated by maximum likelihood using the program BASEML, under the clock, on the species phylogeny, and averaged over the 100 replicates.

Table 2. Posterior means, 95% CIs, and relative errors of divergence times estimates (in My) and molecular rate for a three-species phylogeny.

Software/Calibrations	N	$\tilde{t}_1$	(95% CI)	$\varepsilon(\tilde{t}_1)$	$\tilde{t}_2$	(95% CI)	$\varepsilon(\tilde{t}_2)$	$\tilde{r}$ ($\times 10^{-3}$)	(95% CI)	$\varepsilon(\tilde{r})$
MCMCTree	10^3	10.00	(8.10, 11.97)	0.0%	4.99	(3.96, 6.06)	–0.2%	1.01	(0.81, 1.23)	1%
$t_1 \sim G(100, 100)$	10^4	10.00	(8.09, 11.97)	0.0%	5.11	(4.05, 6.21)	2.2%	1.03	(0.83, 1.24)	3%
$t_2 \mid t_1 \sim \mathrm{U}(0, t_1)$	10^5	10.02	(8.11, 11.99)	0.2%	5.90	(4.71, 7.14)	18.0%	1.21	(0.97, 1.45)	21%
	10^6	10.19	(8.28, 12.16)	1.9%	8.43	(6.82, 10.09)	68.6%	2.91	(2.36, 3.49)	191%
MCMCTree	10^3	10.25	(7.89, 12.47)	2.5%	5.08	(4.03, 6.02)	1.6%	1.00	(0.79, 1.25)	0%
$t_1 \sim \mathrm{B}(0.7, 1.4)$	10^4	10.01	(7.72, 12.16)	0.1%	5.08	(4.03, 6.02)	1.6%	1.04	(0.82, 1.30)	4%
$t_2 \sim \mathrm{B}(0.4, 0.6)$	10^5	8.84	(7.02, 10.51)	–11.6%	5.16	(4.17, 6.04)	3.2%	1.38	(1.12, 1.69)	38%
	10^6	7.10	(6.51, 7.64)	–29.0%	5.78	(5.34, 6.15)	15.6%	4.17	(3.85, 4.53)	317%
BPP	10^3	10.00	(8.08, 11.98)	0.0%	4.98	(3.94, 6.06)	–0.4%	1.01	(0.81, 1.22)	1%
$t_1 \sim G(100, 100)$	10^4	10.00	(8.09, 11.98)	0.0%	5.02	(3.93, 6.15)	0.4%	1.01	(0.81, 1.22)	1%
	10^5	10.00	(8.07, 11.97)	0.0%	5.10	(3.67, 6.61)	2.0%	1.02	(0.80, 1.25)	2%
	10^6	10.00	(8.09, 11.98)	0.0%	5.23	(2.43, 8.12)	4.6%	1.00	(0.75, 1.27)	0%

Note.− The true values are $t_1 = 10$ Ma, $t_2 = 5$ Ma and $r = 10^{-3}$ s/s/My. Posterior means and 95% CIs are averaged across 100 replicate analyses. The τ (distance) estimates from BPP were translated into absolute geological times by sampling from $t_1 \sim G(100, 100)$.

Table 3. Posterior means of divergence times and molecular rate and their relative errors for the nine species phylogenies for various population sizes.

	N	$\tilde{t}_{10}$ (error)	$\tilde{t}_{11}$ (error)	$\tilde{t}_{12}$ (error)	$\tilde{t}_{13}$ (error)	$\tilde{t}_{14}$ (error)	$\tilde{t}_{15}$ (error)	$\tilde{t}_{16}$ (error)	$\tilde{t}_{17}$ (error)	$\tilde{r}$ (error)	P (%)
Young Phylogeny		$\boldsymbol{t_{10} = 10}$	$\boldsymbol{t_{11} = 7}$	$\boldsymbol{t_{12} = 2}$	$\boldsymbol{t_{13} = 4}$	$\boldsymbol{t_{14} = 1}$	$\boldsymbol{t_{15} = 8}$	$\boldsymbol{t_{16} = 3}$	$\boldsymbol{t_{17} = 1}$	$\boldsymbol{r = 1}$	
	10^3	10.6 (6.0%)	7.4 (5.7%)	2.1 (5.0%)	4.2 (5.0%)	1.1 (10.0%)	8.5 (6.3%)	3.2 (6.7%)	1.1 (10.0%)	0.97 (–3.0%)	0
	10^4	10.1 (1.0%)	7.2 (2.9%)	2.2 (10.0%)	4.2 (5.0%)	1.2 (20.0%)	8.2 (2.5%)	3.2 (6.7%)	1.2 (20.0%)	1.02 (2.0%)	0
	10^5	8.0 (–20.0%)	6.0 (-14.3%)	2.5 (25.0%)	4.0 (0.0%)	1.9 (90.0%)	6.6 (-17.5%)	3.3 (10.0%)	1.9 (90.0%)	1.55 (55.0%)	69
	10^6	4.8 (–52.0%)	4.3 (-38.6%)	2.9 (45.0%)	3.6 (-10.0%)	2.7 (170.0%)	4.3 (-46.3%)	3.4 (13.3%)	2.6 (160.0%)	7.28 (628.0%)	100
	Variable	7.4 (–26.0%)	4.9 (-30.0%)	4.1 (105.0%)	2.8 (-30.0%)	0.7 (-30.0%)	5.1 (-36.3%)	2.0 (-33.3%)	1.5 (50.0%)	1.62 (62.0%)	64
Old Phylogeny		$\boldsymbol{t_{10} = 100}$	$\boldsymbol{t_{11} = 70}$	$\boldsymbol{t_{12} = 20}$	$\boldsymbol{t_{13} = 40}$	$\boldsymbol{t_{14} = 10}$	$\boldsymbol{t_{15} = 80}$	$\boldsymbol{t_{15} = 30}$	$\boldsymbol{t_{16} = 10}$	$\boldsymbol{r = 1}$	
	10^3	106.6 (6.6%)	74.7 (6.7%)	21.3 (6.5%)	42.6 (6.5%)	10.7 (7.0%)	85.4 (6.8%)	32.0 (6.7%)	10.7 (7.0%)	0.96 (–4.0%)	0
	10^4	106.5 (6.5%)	74.4 (6.3%)	21.4 (7.0%)	42.6 (6.5%)	10.8 (8.0%)	85.5 (6.9%)	32.2 (7.3%)	10.9 (9.0%)	0.96 (–4.0%)	0
	10^5	103.8 (3.8%)	73.1 (4.4%)	22.3 (11.5%)	42.6 (6.5%)	12.2 (22.0%)	83.4 (4.3%)	32.5 (8.3%)	12.2 (22.0%)	1.00 (0.0%)	0
	10^6	79.5 (-20.5%)	59.4 (-15.1%)	26.0 (30.0%)	39.4 (-1.5%)	19.3 (93.0%)	66.3 (-17.1%)	33.2 (10.7%)	19.1 (91.0%)	1.54 (54.0%)	68
	Variable	78.6 (-21.4%)	54.7 (-21.9%)	29.3 (46.5%)	31.1 (-22.3%)	7.7 (-23.0%)	61.7 (-22.9%)	23.1 (-23.0%)	9.3 (-7.0%)	1.30 (30.0%)	6

Note.− Time estimates are in My and rate estimates in $\times 10^{-3}$ s/s/My. First row (in bold) in each phylogeny denotes the true node ages and rate. Variable means that N varies among lineages as described in Figure 5. P is the percentage of the gene trees that do not match the species tree averaged across all replicates.

Table 4. Posterior means and 95% CIs of divergence times, rate and population sizes for the hominoid phylogeny.

	MCMCTREE	**BPP**	**BPP**		**BPP**
		t_{HC} ~ U(5.7, 10)	r ~ G(100, 200)		No calibration
t_{HCGO}	22.9 (16.3, 28.0)	26.6 (19.5, 33.4)	27.8 (22.6, 33.5)	τ_{HCGO}	13.7 (13.6, 13.9)
t_{HCG}	10.9 (7.8, 13.4)	12.8 (9.5, 16.2)	13.4 (10.9, 16.2)	τ_{HCG}	6.6 (6.6, 6.7)
t_{HC}	8.3 (5.9, 10.1)	7.9 (5.8, 9.9)	8.2 (6.6, 9.9)	τ_{HC}	4.1 (4.0, 4.2)
r	0.80 (0.63, 1.08)	0.53 (0.41, 0.69)	0.50 (0.40, 0.60)	-	-
N_{HCGO}	-	197 (144, 247)	205 (165, 247)	θ_{HCGO}	8.1 (7.8, 8.4)
N_{HCG}	-	85 (62, 106)	88 (71, 106)	θ_{HCG}	3.5 (3.4, 3.6)
N_{HC}	-	147 (106, 190)	154 (123, 186)	θ_{HC}	6.1 (5.7, 6.5)

Note.− Estimates are the posterior means and 95% CIs (in brackets). Divergence times are in My. The rate r is in 10^{-3} s/s/My. The θ (= $4Nrg$) and τ (= rt) parameters are scaled by 10^3. The population sizes, N, are in 10^3. To calculate N, a generation time of 20 years was assumed.

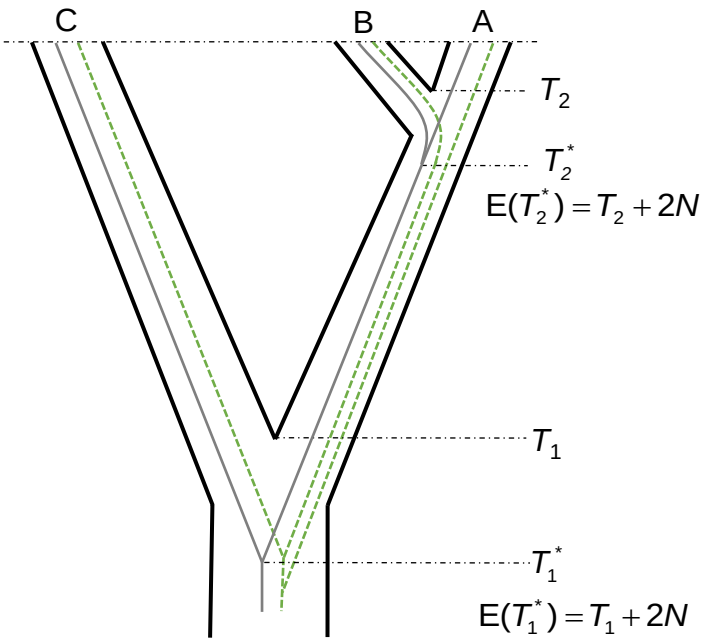

Fig. 1

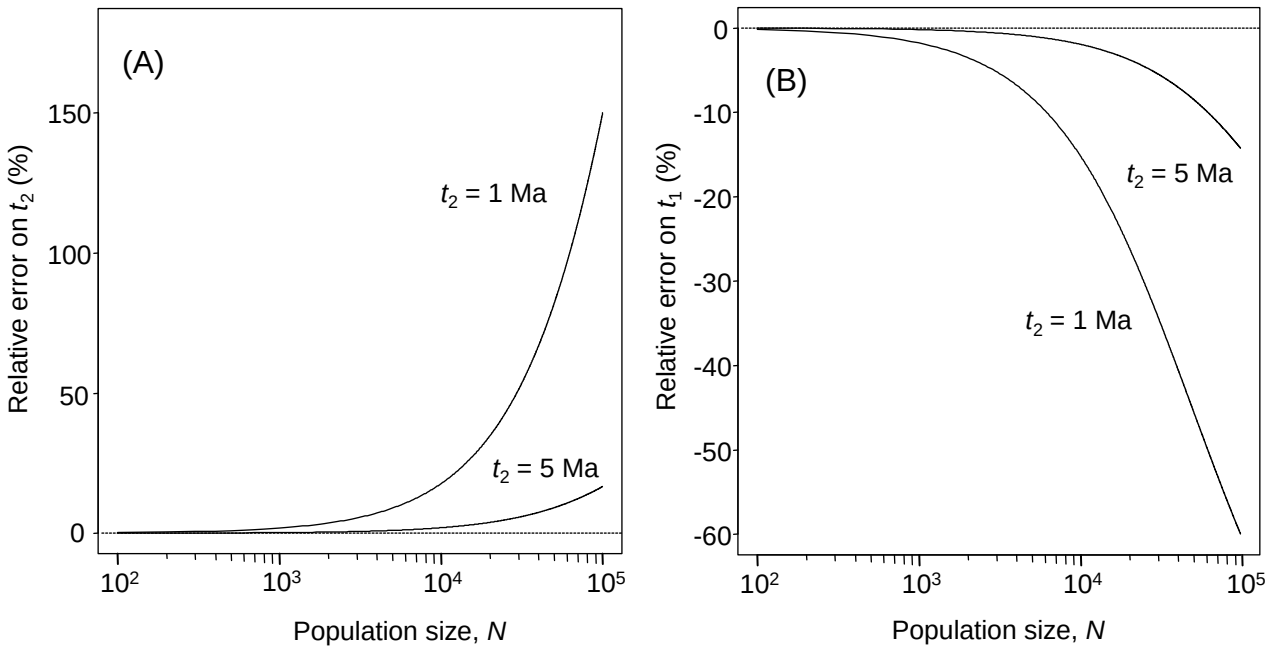
(A)
150
100
50
0
Relative error on t_2 (%)
t_2 = 1 Ma
t_2 = 5 Ma
10^2
10^3
10^4
10^5
Population size, N
(B)
0
-10
-20
-30
-40
-50
-60
Relative error on t_1 (%)
t_2 = 5 Ma
t_2 = 1 Ma
10^2
10^3
10^4
10^5
Population size, N

Fig. 2

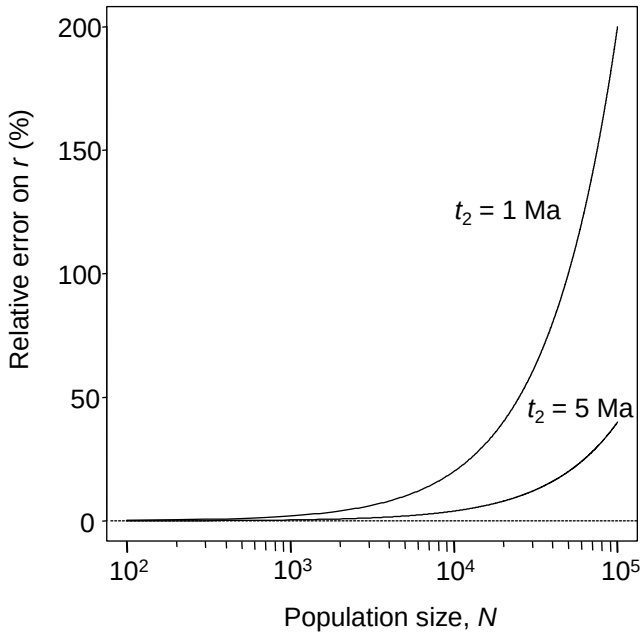
200
150
100
50
0
Relative error on r (%)
t2 = 1 Ma
t2 = 5 Ma
10^2
10^3
10^4
10^5
Population size, N

Fig. 3

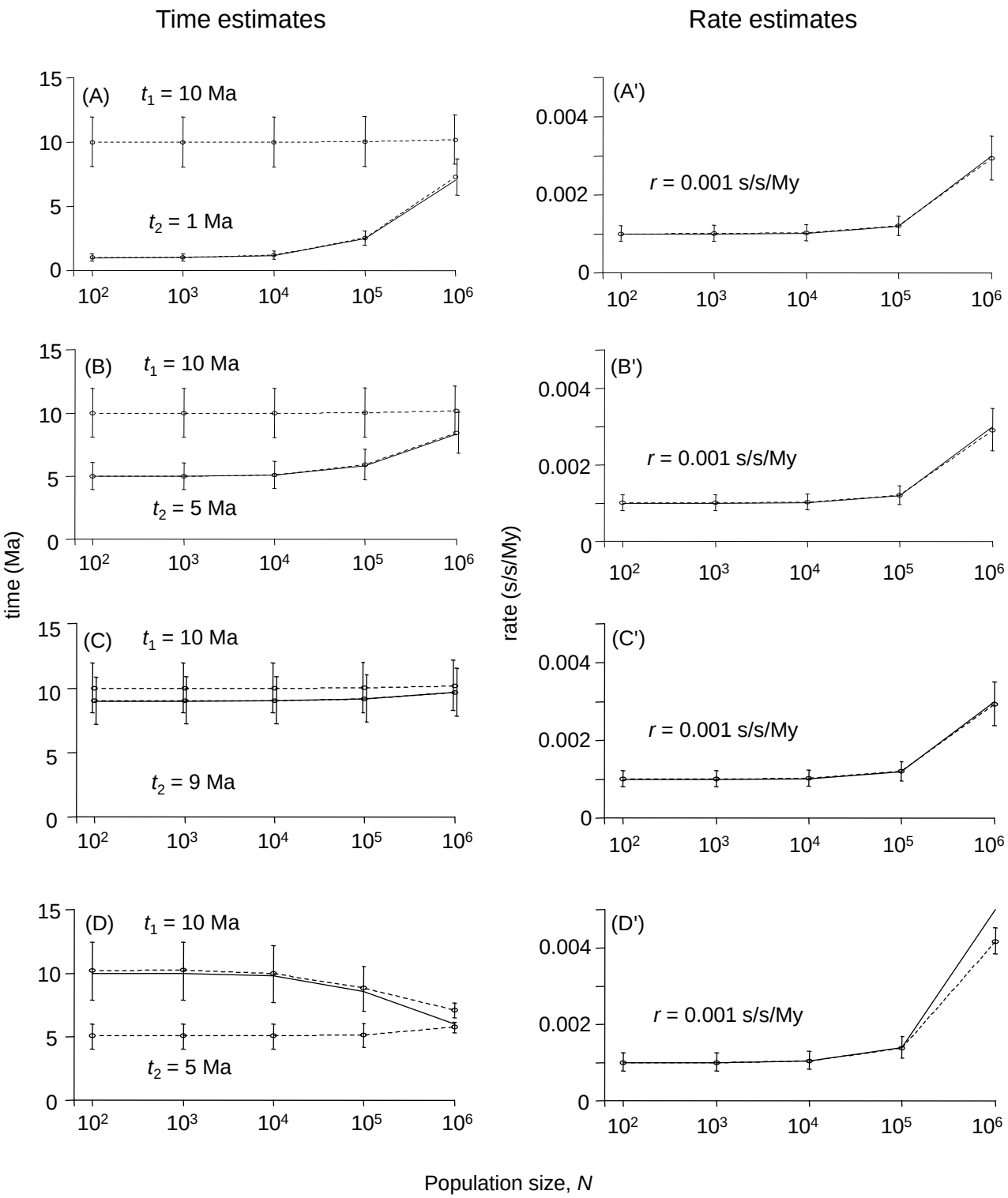
Time estimates
Rate estimates
(A)
t_1 = 10 Ma
t_2 = 1 Ma
(A')
r = 0.001 s/s/My
(B)
t_1 = 10 Ma
t_2 = 5 Ma
(B')
r = 0.001 s/s/My
(C)
t_1 = 10 Ma
t_2 = 9 Ma
(C')
r = 0.001 s/s/My
(D)
t_1 = 10 Ma
t_2 = 5 Ma
(D')
r = 0.001 s/s/My
time (Ma)
rate (s/s/My)
Population size, N

Fig. 4

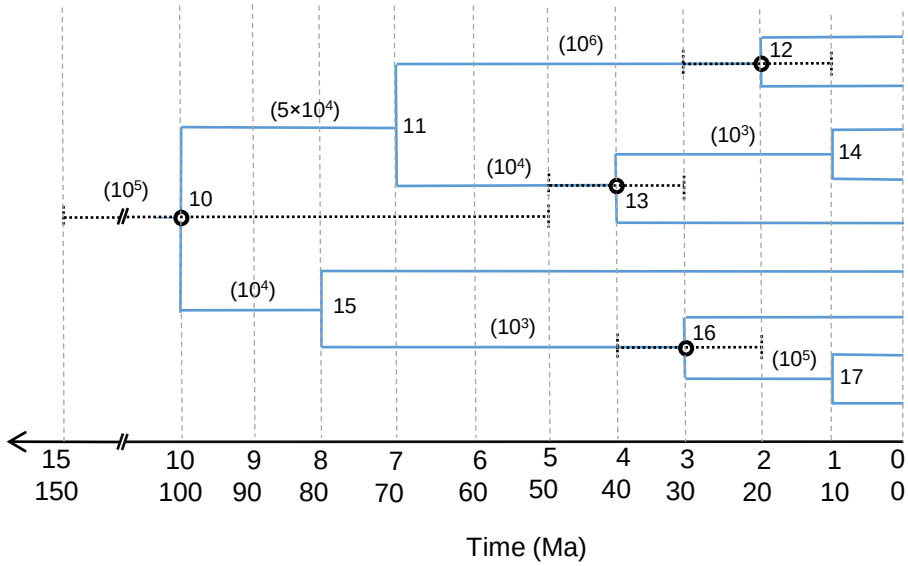
(10⁶)
12
(5×10⁴)
11
(10³)
14
(10⁴)
13
(10⁵)
10
(10⁴)
15
(10³)
16
(10⁵)
17
15 10 9 8 7 6 5 4 3 2 1 0
150 100 90 80 70 60 50 40 30 20 10 0
Time (Ma)

Fig. 5

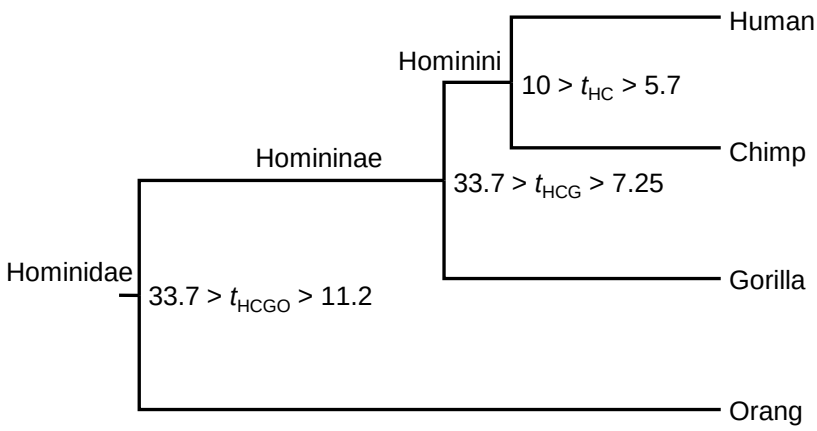
Human
Hominini
10 > t_{HC} > 5.7
Chimp
Homininae
33.7 > t_{HCG} > 7.25
Hominidae
33.7 > t_{HCGO} > 11.2
Gorilla
Orang

Fig. 6