Estimating the Rate of Adaptive Molecular Evolution When the Evolutionary Divergence Between Species is Small

Peter D. Keightley · Adam Eyre-Walker

Received: 23 November 2011 / Accepted: 25 January 2012 © Springer Science+Business Media, LLC 2012

Abstract We investigate the extent by which the estimates of the rate of adaptive molecular evolution obtained by extending the McDonald–Kreitman test are biased if the species, subjected to analysis, diverged recently. We show that estimates can be biased if the nucleotide divergence between the species is low relative to within species variation, and that the magnitude of the bias depends on the rate of adaptive evolution and the distribution of fitness effects of new mutations. Bias appears to be because of three factors: (1) misattribution of polymorphism to divergence; (2) the contribution of ancestral polymorphism to divergence; and (3) different rates of fixation of neutral and advantageous mutations. If there is little adaptive molecular evolution, then slightly deleterious mutations inflate estimates of the rate of adaptive evolution, because these contribute proportionately more to polymorphism than to nucleotide divergence than neutral mutations. However, if there is substantial adaptive evolution, polymorphism contributing to apparent divergence may downwardly bias estimates. We propose a simple method for correcting the different contributions of slightly deleterious and neutral mutations to polymorphism and divergence, and apply it to datasets from several species. We find that estimates of the rate of adaptive molecular evolution from closely related species may be underestimates by ~10% or more. However, after the contribution of polymorphism to divergence is removed, the rate of adaptive evolution may still be overestimated as a consequence of ancestral polymorphism and time for fixation effects. This bias may be substantial if branch lengths are less than 10N_e generations.

Keywords Adaptive evolution · McDonald–Kreitman test · Nucleotide diversity

Introduction

What proportions of nucleotide changes in the genome are neutral and positively selected? What kinds of functional elements experience the highest rates of adaptive evolution? Progress in answering these questions has accelerated in recent years, with the acquisition of large nucleotide sequence datasets of multiple genes, and now with the complete genome sequences of multiple individuals. These datasets enable tests for excess of nucleotide substitutions between species, which can be attributed to adaptive evolution. Fay et al. (2001) (FWW2001) proposed an extension of the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) to estimate \( \alpha \), the fraction of nucleotide substitutions driven to fixation by positive selection (see also Charlesworth 1994). Like all the MK-based methods, FWW2001 assumes that there is a class of sites that evolve under the neutral evolutionary process and a selectively evolving class of sites. At the selectively evolving sites, new mutations are assumed either to be neutral, strongly deleterious or strongly advantageous. Under this model, only neutral mutations contribute to polymorphism, whereas both neutral and advantageous mutations contribute to nucleotide divergence. FWW2001’s estimator for \( \alpha \) is...
\[ z = 1 - \frac{d_s p_n}{d_n p_s}, \]  

where \( d_s \) and \( d_n \) are the numbers of neutral and selected differences per site, respectively, and \( p_s \) and \( p_n \) are the numbers of neutral and selected polymorphisms per site, respectively (note, to maintain consistency with previous articles and accepted nomenclature, we use the subscripts \( n \) and \( s \) for selected and neutral sites, respectively, since these are often used to denote nonsynonymous and synonymous sites). Formula (1) is known to underestimate \( z \) if there are slightly deleterious mutations, since these violate the assumptions of the model by inflating polymorphism relative to divergence. In an attempt to correct this effect, many applications of (1) have excluded polymorphisms in both the neutral and the selected site classes at frequencies below an arbitrary threshold, since low-frequency polymorphisms in the selected class are enriched for deleterious alleles (Fay et al. 2001; Bierne and Eyre-Walker 2004; Andolfatto 2005; Charlesworth and Eyre-Walker 2006; Bachtrog 2008; Haddrill et al. 2008; Ingvarsson 2010). However, this ad hoc procedure also tends to produce biased estimates of \( z \) (Charlesworth and Eyre-Walker 2008).

Recently, ways to explicitly account for the differential contribution of slightly deleterious mutations to polymorphism and divergence have been proposed (Boyko et al. 2008; Eyre-Walker and Keightley 2009 (EWK2009)). EWK2009 suggested integrating over the distribution of fitness effects, \( s \), of new deleterious mutations (the DFE), estimated from polymorphism data, to compute the average fixation probability of a deleterious mutation at the selected sites, \( u_{\text{del}} \). This can then be used to compute the expected divergence at selected sites because of the fixation of slightly deleterious mutations. If the observed divergence exceeds the expected, then the excess is attributed to adaptive substitutions. EWK2009’s estimator for \( z \) is

\[ z = 1 - \frac{d_s u_{\text{del}}}{d_n u_{\text{neut}}}, \]  

where \( u_{\text{neut}} \) is the fixation probability of a neutral mutation. This is therefore a potential improvement over FWW2001, since it attempts to model the contribution of slightly deleterious mutations. However, there are three additional factors that could potentially affect estimates from both FWW2001 and EWK2009:

1. Polymorphisms may be misattributed as divergence in a population sample. Typically, nucleotide divergence is calculated by randomly selecting one allele from the focal species and comparing it to one allele sequenced in the outgroup species (which will usually be a randomly selected allele). As a consequence, some apparent differences between the sequences sampled from the two species will be due to polymorphisms, but not substitutions. Note that the problem is reduced, but not eliminated by using all available alleles, rather than randomly choosing one allele, because there is a lesser chance that a polymorphism will appear to be fixed in a sample of sequences than appearing in a single sequence. However, polymorphism data are usually available for only one species. Polymorphism contributing to apparent divergence potentially has two effects. First, polymorphism tends to dilute the signal of adaptive evolution by inflating estimates of \( d_s \) and \( d_n \). For example, if the nucleotide divergence is low relative to diversity, then many apparent differences at selected sites will be due to effectively neutral polymorphisms, making estimates of adaptive evolution lower than expected. Second, slightly deleterious mutations are expected to contribute proportionally more to polymorphism than divergence and neutral mutations. Polymorphisms contributing to apparent divergence will therefore inflate \( d_n \) proportionally more than \( d_s \), resulting in an overestimation of the rate of adaptive evolution. Polymorphism contributing to divergence is expected to similarly affect estimates of \( \omega_A \), the rate of adaptive substitution scaled by the rate of neutral substitution (Gossmann et al. 2010).

2. A deleterious mutation segregating at the time of the species’ divergence is subsequently more likely to be segregating in one lineage and lost in the other than a neutral mutation.

3. Advantageous mutations go to fixation more rapidly than neutral mutations, and will tend to inflate estimates of adaptive evolution in the short term (Bierne and Eyre-Walker 2004).

We quantify the effects of all of these factors on estimates of the rate of adaptive molecular evolution, and then correct for the first factor by deriving expressions for numbers of neutral and deleterious polymorphisms misattributed as divergence by fitting a distribution of selective effects (\( s \)) to the polymorphism data. Sawyer and Hartl (1992) have proposed a similar method in which the contribution of polymorphism to divergence is corrected using the average selective effect of a new mutation (see also Welch 2006).

**Methods**

**Monte Carlo Simulations**

To investigate the extent by which estimates of the rate of adaptive evolution are affected by the divergence between species, we performed several sets of Monte Carlo simulations. We simulated repeat instances of a new mutation in a population of \( N = 100 \) diploids. There was an ancestral
lineage phase of \( t_1 \) generations, at which point the population
was replicated into two lineages that evolved independently for \( t_2 \) generations. The time \( t_1 \) was set at
\( 20N \) generations, which ensured that the mutation–selection–drift process was close to steady state at the point of
divergence. A new mutation occurred with the probability of
\( p_1 = t_1/(t_1 + 2t_2) \) in the ancestral lineage at a random
generation between 0 and \( t_1 \) generations, otherwise the
mutation occurred in either of the independently evolving
lineages at a random generation \( t_2 \). Individuals were
selected for mating with probability proportional to their
relative fitness. We typically simulated \( 10^7 \) mutations, the
fates of which were tracked to determine the number of
fixation events between the two lineages. We distinguished
between fixations in the independently evolving lineages,
and those originating from a polymorphism present when
the species split that went to fixation in only one lineage (if
the mutation went to fixation in both lineages it was not
counted as a fixation). We also estimated the contribution
of polymorphism to divergence by calculating the proba-
bility that randomly chosen alleles would differ between
the populations if a polymorphism was segregating in one
or both populations at time \( t_2 \).

We assumed that there are two kinds of sites subjected to
the same rates of mutation per site, one kind subjected to
selection and a second evolving neutrally. At the selected
sites, deleterious mutations have additive effects \( (s_{del}, \)
the difference in fitness between the homozygotes) sampled
from a gamma distribution with scale and shape parameters,
\( a \) and \( b \), respectively. Selected mutations had positive fitness
effects \( (s_{adv}) \) with probability \( p_a \) or negative effects with
probability \( 1 - p_a \). In order to parameterise the simulation,
we related \( p_a \) and the mean fixation probability of a new
advantageous or deleterious mutation in a diploid popula-
tion of effective and the actual size \( N, u(N, s) \) (Fisher 1930),
to the proportion of adaptive substitutions \( z \) as follows:
\[
z = \frac{p_a u(N, s_{adv})}{p_a u(N, s_{adv}) + (1 - p_a) u(N, s_{del})},
\]
Rearranging Eq. 1 yields an expression for \( p_a \) in terms
of \( z \) and \( u(N, s_{adv}) \) and \( u(N, s_{del}) \), which depends on
the parameters of the distribution of fitness effects of new
mutations:
\[
p_a = \frac{1}{u(N, s_{adv})/(1/z - 1) + 1}.
\]

Simulations to Investigate the Effect of Polymorphism
on Apparent Divergence

We also performed simulations in which ancestral pol-
ymorphism was excluded, and in which the population was
at mutation–selection–drift equilibrium to exclude the
influence of different rates of fixation of advantageous and
neutral alleles in the initial generations after the split
between the species. We sampled the number of fixed
neutral differences from a binomial distribution for \( x_s \)
sites, assuming that the probability of a substitution at a site is \( d_s \).
The number of fixed deleterious mutations was sampled from
a binomial distribution for \( x_d \) sites, assuming that the probability
of a substitution at a site is \( d_d \), where \( d_d \) is the
average fixation probability of a new deleterious
mutation (Fisher 1930). We then incorporated a contribu-
tion to apparent divergence from segregating polymor-
phism. We used transition matrix methods (Keightley and
Eyre-Walker 2007) to generate vectors representing the
equilibrium frequency distributions of alleles originating
from new mutations, denoted as \( v(s) \) and \( v(n) \) for neutral
and selected mutations, respectively, for a model diploid
population of constant size \( N \) and an arbitrary mutation
rate. Let \( \theta_s \) be the neutral nucleotide diversity in the model
population, i.e. the probability that two alleles are different
in state if randomly sampled without replacement:
\[
\theta_s = \sum_{i=1}^{2N-1} \frac{i}{2N} \left( \frac{1}{2N-1} - \frac{i}{2N} \right) + \left(1 - \frac{i}{2N}\right) \left( \frac{i}{2N-1} \right) \nu(s).
\]
(5)
For an actual neutral diversity, \( \theta_n \), which is assumed as a
parameter of the simulation model, the expected number of
segregating neutral sites is then
\[
n_{segs} = x_s (\theta_s/\theta_s') \sum_{i=1}^{2N-1} \nu(s),
\]
(6)
Similarly, the expected number of segregating selected
sites is
\[
n_{segn} = x_n (\theta_s/\theta_s') \sum_{i=1}^{2N-1} \nu(n),
\]
(7)
We sampled the contribution of polymorphic sites to
nucleotide divergence, under the assumption that one allele
is sampled at random from each of the two lineages related
by a common ancestor as follows. For each of \( 2n_{segs} \)
segregating neutral sites (the number of sites is multiplied by
two to account for the contributions from polymorphism in
the two lineages), we sample a number of derived alleles, \( y \),
from \( v(s) \) in proportion to the frequency of each element of
the vector. The frequency of the derived allele sampled is
then \( q = y/(2N) \). This segregating site contributes one
apparent difference with probability \( q \). In the same way,
we sampled the number of polymorphic selected sites
contributing to apparent divergence by sampling from \( v(n) \).
We also used the transition matrix approach to generate site
frequency spectra (SFSs) for neutral and selected sites.
These are the distributions of numbers of derived (i.e., mutant) alleles in a sample of \(x_s\) and \(x_n\) neutral and selected sites, respectively.

**Analysis of Data**

The data from the simulations consisted of SFSs for neutral and selected sites and the number of selected and neutral differences, including apparent differences attributable to polymorphism, and corresponding numbers of sites. The first stage of the analysis was to perform the inference procedure described by Keightley and Eyre-Walker (2007), which estimates the parameters of the DFE using the SFSs assuming a gamma distribution of mutational effects. We then produced uncorrected estimates of \( \alpha \) (or \( \omega_A \)) using estimated parameters of the DFE to predict the number of fixed differences in the absence of adaptive evolution (EWK2009). Corrected estimates of \( \alpha \) (or \( \omega_A \)) were then generated by estimating the numbers of differences attributable to polymorphism, as described below, and subtracting these numbers from the observed numbers of differences, using the method of EWK2009.

**Results**

Estimating the rate of adaptive evolution within the MK framework, is likely to be complicated if the divergence between species is small relative to the diversity within species. To investigate the overall extent of bias originating from short branch length, we performed Monte Carlo simulations in which a species was split into two daughter species, which were then allowed to evolve independently. We then used the method of EWK2009 to estimate \( \alpha \), the proportion of adaptive substitutions. Our simulation results suggest that \( \alpha \) is overestimated if the time since speciation is short (Fig. 1). However, as the time since divergence increases, this effect dissipates at a rate that depends on the rate of adaptive evolution. Apparently, the higher the rate of adaptive evolution, the faster the effect disappears. In the following sections, we attempt to tease apart the causes of these biases.

**Polymorphism Contributing to Apparent Nucleotide Divergence**

One potential cause for the overestimation of \( \alpha \) for small divergence between species is the misattribution of polymorphism to divergence. To investigate this, we performed simulations based on transition matrix methods, assuming an equilibrium model and no contribution from ancestral polymorphism. Figure 2 shows estimates of the proportion of adaptive substitutions, \( \alpha \), for four simulated \( \alpha \) values plotted against \( \tau_S = d_s/\theta_S \), the ratio of divergence to diversity at neutral sites. Results are presented for two distributions of fitness effects of new mutations: an exponential distribution (\( b = 1 \); Fig. 2a) and a strongly leptokurtic distribution (\( b = 0.1 \); Fig. 2b). In Fig. 2a most mutations are strongly deleterious (\( Ns \ll -1 \)), but there are more slightly deleterious mutations (\( Ns \sim -1 \)) than in Fig. 2b. In Fig. 2b, most mutations are either nearly neutral (\( Ns \rightarrow 0 \)) or strongly deleterious. As expected, if the level of divergence is high relative to diversity (i.e., \( \tau_S > 50 \)), estimates of \( \alpha \) are always essentially unbiased. However, for lower values of \( \tau_S \), the estimated level of adaptive evolution can either be under- or over-estimated, depending on the true level of adaptive evolution and the proportion of effectively neutral (\( Insl < 1 \)) and slightly deleterious (\( 1 < Ns < 10 \)) mutations. There are two processes acting in opposite directions that explain these results. The presence of slightly deleterious mutations tends to bias estimates of \( \alpha \) upward, because these mutations contribute proportionately more to polymorphism than divergence relative to neutral mutations. This results in a higher contribution of polymorphism to apparent divergence for the selected sites than the neutral sites, which is obvious when there is no adaptive evolution (\( \alpha = 0 \)). Effectively neutral mutations (\( Insl < 1 \)), however, have the opposite effect. If there is some adaptive evolution, these mutations dilute the signal of adaptive evolution.
if the branch lengths are short. For example, consider the case where there are only adaptive and neutral mutations at selected sites. If divergence is small relative to diversity, the apparent divergence is dominated by neutral polymorphisms. In the extreme case, the ratio of apparent divergence to diversity is the same as that for neutral sites.

A Simple Correction for the Contribution of Polymorphism to Apparent Divergence

The contribution of polymorphism to apparent divergence can easily be corrected by using the estimate of the DFE based on the distribution of allele frequencies in the population sample as follows. From the observed SFSs for neutral and selected sites in the focal species, we use the method of Keightley and Eyre-Walker (2007) to obtain estimates of parameters of the distribution of fitness effects of new mutations, along with corresponding estimates of $v(s)$ and $v(n)$ for a population of $N$ diploids. The number of neutral polymorphic sites contributing to nucleotide divergence is computed from

$$d'_s = 2xs \sum_{i=1}^{2N-1} \frac{i}{2N} v(s)_i. \quad (8)$$

Similarly, the number of apparently fixed selected sites is computed from (8), with $s$ replaced by $n$. In order to infer corrected estimates of $x$ or $\omega_\Lambda$, the number of apparently fixed neutral and selected sites are subtracted for the corresponding observed number of differences.

We used this procedure to correct EWK2009 estimates of $x$ and $\omega_\Lambda$ in simulations described above by removing estimates of the numbers of apparent differences attributable to polymorphism. We found that mean corrected estimates closely match the true values for $x$ and $\omega_\Lambda$ in all cases investigated (data not shown).

Residual Effects of Short Branch Length

We then used Monte Carlo simulations to investigate the extent by which estimates of rates of adaptive substitution are affected by polymorphism segregating in the ancestral population before the species divergence and by effects attributable to differences in the rates, at which selected and neutral mutations go to fixation. Figure 3 shows estimates of $x$, corrected for polymorphism contributing to divergence, so any remaining bias is attributable to ancestral polymorphism and time to fixation effects. The results suggest that $x$ tends to be overestimated in all these cases, particularly when there are relatively large numbers of slightly deleterious mutations (Fig. 3a). There are two reasons for this: First, for alleles segregating at the time at which the two species split, conditional on the derived (i.e., mutant) allele remaining segregated in one lineage, a deleterious mutation has a higher probability than a neutral mutation of becoming lost in the second lineage, and therefore contributing to nucleotide divergence. As a consequence of this effect, ancestral polymorphism originating from slightly deleterious mutations contribute proportionally more to divergence than neutral mutations, and inflate estimates of $x$ in the generations immediately after the speciation event. Second, selected mutations fix more rapidly than neutral mutations, so contribute proportionally more to fixed differences in the generations immediately after the species split than at equilibrium. These effects are illustrated in Fig. 4, which shows the numbers of ancestral polymorphisms and new mutations that contribute to the true divergence (i.e., fixed differences between species), and apparent divergence (i.e., polymorphisms that appear as substitutions). It is evident that slightly deleterious mutations segregating at the time of
divergence (green line in Fig. 4b) contribute substantially more apparent fixations just after species divergence than new mutations (red line in Fig. 4b). It is also evident that advantageous mutations fix more rapidly than neutral mutations, i.e., the blue line starts to climb steeply sooner for advantageous mutations (Fig. 4c) than neutral mutations (Fig. 4a). Overall, however, the results (Fig. 3) suggest that the effect of branch length becomes quite small by $10N$ generations, and has virtually disappeared by $20N$ generations. Note that in Fig. 3, $x$ tends to be underestimated if the true value of $x = 0.5$ for $t \gg 20N$ generations if $N_{s_{adv}}$ is small. This is due to the contribution of slightly advantageous alleles to the SFS, an affect that we attempt to account for elsewhere (Schneider et al. 2011).

Discussion

In this article, we have attempted to quantify the effects of short branch lengths on estimates of the rate of adaptive molecular evolution. In cases where there is substantial neutral site diversity relative to divergence (i.e., $t_S = d_S/\theta_S < 10$), polymorphism misattributed to divergence can have a significant impact on estimates of the fraction of adaptive substitutions ($x$) and the relative rate of adaptive substitution ($\omega_A$). However, the direction of the bias depends on the true value of $x$ (or $\omega_A$) and on the parameters of the distribution of effects of new mutations, which determines the frequency at which fixations and polymorphisms originate from slightly deleterious mutations. In general, if $x \to 0$, estimates of $x$ are expected to be over-estimated, because slightly deleterious mutations contribute proportionately more to diversity than to divergence and neutral mutations. However, if the true value of $x > 0$, $x$ can potentially be underestimated, particularly if few slightly deleterious mutations contribute to nucleotide divergence (Figs. 1, 2). We have suggested a simple method to correct estimates of $x$ and $\omega_A$ (based on Eq. 8) by estimating the number of segregating polymorphisms contributing to divergence. What effect does this correction have on estimates of $x$ from real data? To investigate this, we have re-analysed a number of published polymorphism datasets for a range of species (Drosophila, primates, and murid rodents), and applied the correction (Table 1). We find that, as expected, only estimates from the analysis of closely related species (i.e., D. melanogaster versus D. simulans and Mus musculus castaneus versus M. famulus), where $t_S < 10$, are substantially affected. In these cases, corrected estimates of $x$ are somewhat higher than uncorrected estimates.

Within the EWK2009 method for estimating the rate of adaptive evolution, divergence is measured by taking a single allele from each species or calculating the mean divergence across the loci. A simpler means of correcting for the contribution of polymorphism to divergence would be to consider all alleles by, for example, using the consensus of these sequences. However, this strategy is not possible if there are polymorphism data for one species only. Furthermore, using the consensus will reduce, but not eliminate the bias associated with polymorphisms contributing to apparent divergence, because a polymorphism may still appear to be fixed in a sample of sequences, and hence appear to be a fixed difference between the two species. If polymorphism data are available only for one species, we need to assume that the contribution is the same in the two species. If polymorphism data from both
species were available, this assumption could be avoided, but such datasets are currently relatively rare. If the level of polymorphism is different in the outgroup species compared to the ingroup species then the rate of adaptive evolution may therefore be underestimated. We have only considered the effect of divergence on estimates of adaptive evolution estimated by the EWK2009 method, since this is expected to yield unbiased estimates if divergences are reasonably large. However, low divergences between species will affect estimates from other methods, such as that of Fay et al. (2001), and Bierne and Eyre-Walker (2004), but the biases will be different, and they will depend on whether there are slightly deleterious mutations and whether low frequency mutations have been removed from the sample. We have also assumed independence between sites, though simulations suggest that estimates are reasonably unbiased unless linkage is tight (EWK2009). Our simulations have also assumed equilibrium demography, but demographic effects can be estimated within the model (Keightley and Eyre-Walker 2007).

Although we have suggested a simple measure to correct the effect of polymorphism on divergence, there are two other sources of bias that are more difficult to correct. First, slightly deleterious mutations segregating in the ancestral population have a greater chance of being lost in one lineage, but continuing to segregate in the second lineage than neutral mutations. Second, advantageous mutations contribute to divergence quicker than neutral alleles. These effects both lead to biased estimates of the rate of adaptive evolution upward (Fig. 3). However, the magnitude of this

Fig. 4 Contributions of neutral (a), deleterious (b), and advantageous mutations (c) to true and apparent nucleotide divergence. Nucleotide differences are classified into four types: fixations resulting from mutations that occurred after the split between the species (true fix), differences caused by polymorphism originating from mutations that occurred after the species split (apparent), fixations due to mutations that segregated at the time of the species split (ancestral), and differences caused by polymorphism originating from mutations that segregated at the time of the species split (ancestral apparent). The simulation parameters were as in Fig. 3a, with $N_{d} = -10$ and $N_{adv} = +10$

Table 1 Uncorrected and corrected (using Eq. 8) estimates of the proportion of adaptive substitutions ($\alpha$) at amino acid replacement sites, obtained from the analysis of polymorphism data from the focal species compared with divergence to the outgroup species

<table>
<thead>
<tr>
<th>Focal species</th>
<th>Outgroup species</th>
<th>References</th>
<th>Neutral sites</th>
<th>Selected sites</th>
<th>$\tau$</th>
<th>Uncorrected $\alpha$</th>
<th>Corrected $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td>D. simulans</td>
<td>Shapiro et al. (2007)</td>
<td>4-folds</td>
<td>0-folds</td>
<td>6.2</td>
<td>0.52</td>
<td>0.57</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Macaca mulatta</td>
<td>Akey et al. (2004)</td>
<td>Intron 0-folds</td>
<td>71</td>
<td>0.13</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>M. musculus castaneus</td>
<td>M. famulus</td>
<td>Halligan et al. (2010)</td>
<td>Intron 0-folds</td>
<td>4.3</td>
<td>0.45</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>M. musculus castaneus</td>
<td>Rattus norvegicus</td>
<td>Halligan et al. (2010)</td>
<td>Intron 0-folds</td>
<td>21</td>
<td>0.37</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>M. musculus castaneus</td>
<td>M. famulus</td>
<td>Kousathanas et al. (2011)</td>
<td>Intron Noncoding sites 5' of genes</td>
<td>4.3</td>
<td>0.12</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>
bias is likely to be relatively small in practice. Consider again the two pairs of most closely related species *D. melanogaster* versus *D. simulans* and *M. m. castaneus* versus *M. famulus*. For the *Drosophila* species, assuming 10 generations per year, a speciation date of 2.5MYA (Powell and DeSalle 1995), and an effective population size of $1.5 \times 10^6$ (Eyre-Walker et al. 2002), the predicted time since divergence is $\sim 17N$ generations. For *M. m. castaneus* versus *M. famulus*, assuming two generations per year, a divergence date of 2.5MYA [assuming that mouse and rat diverged 12MYA (Michaux et al. 2001; Benton and Donoghue 2007) and a local molecular clock], and an effective population size for *M. m. castaneus* of $5.8 \times 10^5$ (Halligan et al. 2010), the predicted time since divergence is $\sim 9N$ generations. In both cases, estimates of $\alpha$ are therefore predicted to be inflated, but only by a small extent due to short branch length and ancestral polymorphism effects (Fig. 3).

Estimating the rate of adaptive evolution is problematic, if the divergence time between two species is short. Bias originating from polymorphism misattributed as divergence can be corrected under the assumption that levels of diversity and the distribution of effects of selected mutations are similar in the two species. However, our proposed correction does not entirely remove the bias, so we recommend that species are chosen such that the branch length of each species to the common ancestor is greater than $10N_e$ generations.

**Acknowledgments** PDK acknowledges the grant support from the BBSRC and Wellcome Trust. The authors thank Dan Halligan for helpful comments in the manuscript.

**References**


