Title: Estimating Selection Intensity on Synonymous Codon Usage in a Non-equilibrium Population

Authors: Kai Zeng*† and Brian Charlesworth*

Affiliations:

* Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

† State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-Sen University, Guangzhou 510275, China
Running head: Selection on Codon Usage

Key words: codon usage bias, demography, natural selection, *Drosophila melanogaster*

Corresponding author:

Kai Zeng

Address: Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King’s Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Telephone: +44 131 650 5543

Email: kai.zeng.cn@gmail.com
Codon usage bias is the nonrandom use of synonymous codons for the same amino acid. Most population genetic models of codon usage evolution assume that the population is at mutation-selection-drift equilibrium. Natural populations, however, frequently deviate from equilibrium, often because of recent demographic changes. Here, we construct a matrix model that includes the effects of a recent change in population size on estimates of selection on preferred versus unpreferred codons. Our results suggest that patterns of synonymous polymorphisms affecting codon usage can be quite erratic after such a change; statistical methods that fail to take demographic effects into account can then give incorrect estimates of important parameters. We propose a new method that can accurately estimate both demographic and codon usage parameters. The method also provides a simple way of testing for the effects of covariates such as gene length and level of gene expression on the intensity of selection, which we apply to a large Drosophila melanogaster polymorphism dataset. Our analyses of two-fold degenerate codons reveal that: (i) selection acts in favor of preferred codons (ii) there is mutational bias in favor of unpreferred codons (iii) shorter genes and genes with higher expression levels are under stronger selection (iv) there is little evidence for a recent change in population size in the Zimbabwe population of D. melanogaster.
Codons specifying the same amino acid are called synonymous codons. These are often used non-randomly, with some codons appearing more frequently than others. This biased usage of synonymous codons has been found in many organisms such as Drosophila, yeast, and bacteria (Ikemura 1985; Duret and Mouchiroud 1999; Hershberg and Petrov 2008). Conventionally, synonymous codons for a given amino acid are divided into two classes: preferred and unpreferred codons (Ikemura 1985; Akashi 1994; Duret and Mouchiroud 1999). Several observations indicate that codon usage is affected by natural selection. First, in species with codon usage bias, preferred codons generally correspond to the most abundant tRNA species (Ikemura 1981). Second, highly expressed genes usually have higher codon usage bias than genes with low expression (Sharp and Li 1986; Duret and Mouchiroud 1999; Hey and Kliman 2002). Third, the synonymous substitution rate of a gene has been shown to be negatively correlated with its degree of codon usage bias (Sharp and Li 1986; Bierne and Eyre-Walker 2006). The most commonly cited explanations of the apparent fitness differences between preferred and unpreferred codons are selection for translation efficiency, translational accuracy, and mRNA stability (Ikemura 1985; Eyre-Walker and Bulmer 1993; Akashi 1994; Drummond et al. 2005). Recently, it has been proposed that exon splicing also affects codon usage bias (Warnecke and Hurst 2007).

From a population genetics perspective, the extent of codon usage bias is ultimately a product of the joint effects of mutation, selection, genetic drift, recombination, and demographic history. The Li-Bulmer model of drift, selection and reversible mutation between preferred and unpreferred codons at a site is the most widely used model (Li 1987;
Bulmer 1991; McVean and Charlesworth 1999). Applications of this model generally assume that the population is at mutation-selection-drift equilibrium. However, empirical studies have suggested that changes in the strengths of various driving forces may not be unusual. For example, in Drosophila melanogaster, there is evidence that the population size (Li and Stephan 2006; Thornton and Andolfatto 2006; Keightley and Eyre-Walker 2007; Stephan and Li 2007), recombinational landscape (Takano-Shimizu 1999), and mutational process (Takano-Shimizu 2001; Kern and Begun 2005) may have changed significantly over the species’ evolutionary history.

Such changes cause departures from equilibrium. Theoretical models show that it takes a very long time, proportional to the reciprocal of the mutation rate, for the population to approach a new equilibrium state (Tachida 2000; Comeron and Kreitman 2002). Before reaching equilibrium, the population often shows counterintuitive patterns of evolution (Eyre-Walker 1997; Takano-Shimizu 1999, 2001; Comeron and Kreitman 2002; Comeron and Guthrie 2005; Charlesworth and Eyre-Walker 2007). Despite these theoretical results, details of the patterns of polymorphism and substitution rates following a recent change in population size, and their effects on estimates of strength of selection, have not been determined.

The above findings point to the importance of incorporating non-equilibrium factors into the study of codon usage bias. To this end, we extend the Li-Bulmer model to allow population size to vary over time, by representing the evolutionary process by a transition matrix. By analyzing this matrix model, we show that a recent change in population size can result in erratic patterns of codon usage, and that methods failing to take into account these
demographic effects can give false estimates of the intensity of selection.

To solve these problems, we propose a new method, which does not require polarizing ancestral versus derived states using outgroup data (c.f. CUTTER and CHARLESWORTH 2006), but only knowledge of preferred versus unpreferred states defined by patterns of codon usage. We use information on both polymorphic and fixed sites, which enables both mutational bias and the strength of selection to be estimated, in contrast to previous methods that use information on polymorphisms alone. Simulations indicate that this method can accurately estimate both demographic and codon usage parameters, and can distinguish between selection and demography. We use the new method to analyze a large D. melanogaster polymorphism dataset (SHAPIRO et al. 2007), and find evidence for natural selection on synonymous codons. We use our approach to show that genes with shorter coding sequences and higher levels of expression are under significantly stronger selection than longer genes with lower expression.
MATERIALS AND METHODS

Description of the model: We consider a diploid Wright-Fisher population (EWENS 2004, p. 21), whose size in generation $t$ is $N_t$. At an autosomal nucleotide site, two variants, $A$ and $a$, can occur. Mutation is reversible: the per generation mutation rate from $a$ to $A$ is $u$, and the rate in the opposite direction is $ku$. The fitnesses of the three possible genotypes, $AA$, $Aa$, and $aa$ are $1$, $1 - \frac{1}{2}s$, and $1 - s$, respectively (the genic selection model). Denote the number of occurrences of allele $A$ in generation $t$ by $X(t)$. If $X(t) = i$, and assuming that mutation precedes selection, the frequency of $A$ after mutation is

$$x^*(t) = \frac{i}{2N_t} (1 - ku) + \left(1 - \frac{i}{2N_t}\right)u. \quad (1)$$

After selection, this becomes

$$x(t) = x^*(t) \left[1 - \frac{1}{2} s \left(1 - x^*(t)\right)\right]. \quad (2)$$

The transition probability for $X(t)$ is thus

$$p_{ij}(t) = \Pr\{X(t+1) = j \mid X(t) = i\} = \binom{2N_{t+1}}{j} \left(x(t)\right)^j (1 - x(t))^{2N_{t+1} - j}. \quad (3)$$

In Equations 1 – 3, we have $0 \leq i \leq 2N_t$ and $0 \leq j \leq 2N_{t+1}$.

Let $f_j(t) = \Pr(X(t) = j)$ be the probability that $A$ is represented $j$ times in the population at time $t$. This satisfies the relation

$$\sum_{j=0}^{2N_t} f_j(t) = 1. \quad (4)$$

Standard Markov chain theory (KARLIN and TAYLOR 1975) implies the following recursion relationship
\[ f_j(t+1) = \sum_{i=0}^{2N_t} f_i(t) p_{ij}(t), \quad 0 \leq j \leq 2N_{t+1}. \] (5)

In an equilibrium population with constant size \( N \), Equation 5 can be written as

\[ f_j = \sum_{i=0}^{2N} f_i p_{ij}, \quad 0 \leq j \leq 2N. \] (6)

Equation 6 is a system of linear equations subject to the requirement given in Equation 4.

Following the method used to calculate the frequency spectrum of polymorphic sites under the infinite sites model (e.g. Equation 16 in McVean and Charlesworth 1999), we can decompose \( f_j(t) \) into two sub-processes

\[ f_j(t) = f_{A,j}(t) + f_{a,2N_t-j}(t), \quad 1 \leq j \leq 2N_t - 1. \] (7)

Here, \( f_{A,j}(t) \) is the probability that the mutant allele \( A \) is currently represented \( j \) times in the population, having originated at a site fixed for \( a \). We call these mutations “\( a \rightarrow A \) polymorphic mutations.” The second sub-process involves mutations originating at sites fixed for \( A \), such that the mutant allele \( a \) is currently represented \( 2N_t - j \) times in the population (i.e. \( A \) is represented \( j \) times); these are \( A \rightarrow a \) polymorphic mutations. For the first sub-process, we have the following recursion formula

\[ f_{A,j}(t+1) = \sum_{i=1}^{2N_t-1} f_{A,i}(t) p_{ij}(t) + f_0(t) p_{0j}(t), \quad 1 \leq j \leq 2N_{t+1} - 1. \] (8)

The first term on the right describes the dynamics of sites that are polymorphic in the \( t \)-th generation, and the second term describes new \( a \rightarrow A \) polymorphic mutations. This formula is analogous to Equation 3 of Evans et al. (2007), which was derived under the infinite sites model. A similar formula holds for the second sub-process. Equations 7 and 8 provide an alternative way of calculating the frequency spectrum of polymorphic sites, on the infinite sites assumption that all new mutations arise at sites that are fixed for either \( A \) or \( a \); but in fact
they apply more generally; some examples are given in supplemental Figure S1. While the application of our model to data analysis does not depend critically on the infinite sites assumption, this is likely to be a good approximation in practice, and is valid for the simulations described in the first part of the Results section.

From the two sub-processes, the proportion of polymorphic sites originating from sites previously fixed for \( a \), \( P_a \), is given by

\[
P_a(t) = \frac{\sum_{j=1}^{2N_t-1} f_{A,j}(t)}{\sum_{j=1}^{2N_t-1} (f_{A,j}(t) + f_{a,j}(t))}.
\]  

(9)

The proportion of polymorphic sites that originated from sites previously fixed for \( A \), \( P_A \), can be calculated similarly. Furthermore, we can calculate the rate at which a site currently fixed for \( a \) is replaced by a site fixed for \( A \), denoted by \( r_{aA}(t) \)

\[
r_{aA}(t) = \sum_{j=1}^{2N_t-1} f_{A,j}(t-1) p_{j,2N_t}(t-1) + f_0(t-1) p_{0,2N_t}(t-1).
\]  

(10)

Similarly, we can calculate the rate, \( r_{Aa}(t) \), at which a site currently fixed for \( A \) is replaced by a site fixed for \( a \). Thus, the total substitution rate is

\[
r(t) = r_{aA}(t) + r_{Aa}(t).
\]  

(11)

In an equilibrium population, Equations 7 – 11 are independent of \( t \).

**Statistical methods:** In this section, we use the theory developed above to construct two new methods which are based on the frequencies of preferred versus unpreferred variants at nucleotide sites, and do not require the use of an outgroup to infer ancestral versus derived states, extending the similar approach of Cutter and Charlesworth (2006). Let \( d_i \) be the number of sites at which allele \( A \) is represented \( i \) times in a sample of \( n \) alleles from a
population. The frequency distribution for the sample, \( d \), is a vector of all the \( d_i \)'s (\( 0 \leq i \leq n \)); this includes cases where the sample is fixed for either \( A \) or \( a \).

We start with the simple case where the population is at equilibrium, denoted as \( L_0 \). The population dynamics are determined by the parameters: \( N, u, \kappa, \) and \( s \). However, if all evolutionary forces are weak so that the diffusion approximation holds, the state of a sample from the population is described by the compound parameters: \( \theta = 4Nu \) and \( \gamma = 2Ns \) (EWENS 2004, Chap. 5). The distribution of the numbers of copies of \( A \) per site in the population can be calculated by solving Equation 6 numerically (see below). We write this distribution as \( f \), which is a vector with \( 2N + 1 \) elements. The log-likelihood for the sample, \( d \), is then

\[
L_0(d \mid \gamma, \theta, \kappa) = \sum_{i=0}^{n} d_i \times \log \left[ \sum_{j=0}^{2N} f_j \left( \begin{array}{c} n \\ i \end{array} \right) \left( \frac{j}{2N} \right)^{\gamma - 1} \left( 1 - \frac{j}{2N} \right)^{\kappa - 1} \right].
\] (12)

In Equation 12, we have adopted the “conventional” assumption that sites behave independently of each other (AKASHI and SCHAEFFER 1997; EYRE-WALKER 1997; McVEAN and VIEIRA 1999; McVEAN and CHARLESWORTH 2000; MASIDE et al. 2004; CUTTER and CHARLESWORTH 2006; GALTIER et al. 2006). Although this assumption is unrealistic, it is mathematically tractable and seems to work fairly well in practice with relatively free recombination (e.g. with a local recombination rate higher than 2cM/Mb, K. ZENG unpublished results; see also WILLIAMSON et al. 2005; and BOYKO et al. 2008).

To model the effects of a change in population size, we assume that the population is originally at equilibrium with population size \( N_b \). Its population size then changes instantly to \( N_a \), and stays constant for \( t \) generations, at which point a sample is taken from the population. Hence, the model, denoted as \( L_1 \), has five parameters: \( \gamma_b \) (\( = 2N_b s \)), \( \theta_b \) (\( = 4N_b u \)), \( \kappa, \) \( g \) (\( = N_a / N_b \)), and \( \tau \) (\( = t / N_a \)). The demographic model underlying \( L_1 \), although somewhat unrealistic, is
mathematically tractable, and has been widely used (Williamson et al. 2005; Li and Stephan 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008).

Let the distribution of the numbers of A per site in the generation just before the change in population size be $f^b$. Then the distribution in the $t$-th generation after the change, $f^a(t)$ (a vector with $2Na + 1$ elements), can be obtained by iterating Equation 5 with $f^b$ as the initial condition. The log-likelihood of the data is now

$$L_1(d | \gamma, \theta, \kappa, g, t) = \sum_{i=0}^{2Na} d_i \times \log \left[ \sum_{j=0}^{n} f^a_j(t) \left( \frac{n}{2Na} \right)^i \left( 1 - \frac{j}{2Na} \right)^{n-i} \right].$$

(13)

Note that, when $g = 1$ and/or $\tau = \infty$, the more complex model, $L_1(d | \gamma, \theta, \kappa, g, t)$, reduces to the simpler model, $L_0(d | \gamma, \theta, \kappa)$. Here, following Williamson et al. (2005), we use a chi-square test with two degrees of freedom to distinguish these two models under the assumption that it will yield a conservative test.

**Numerical computations:** We used standard numerical methods to solve the linear system given in Equation 6: an LU-decomposition procedure followed by one round of numerical improvement (Press et al. 1992). The adequacy of this numerical method was checked by comparing the results with those obtained by iterating Equation 5 (results not shown). To find the maximum likelihood estimates (MLEs) of the parameters, we used the simplex algorithm (Press et al. 1992). Multiple start points were used to make sure that the global maximum was found. To save computer time, we used a relatively loose search criterion in the simulations, with a sparse grid of values and a relaxed convergence criterion.

When implementing the model, the size of the transition matrix has to be specified. As mentioned above, population dynamics in the diffusion limit are determined by the
compound parameters $\theta$ and $\gamma$. Hence it is legitimate to “scale down” the population size to give a tractable size of matrix, preserving the $\theta$ and $\gamma$ values for the true population size. This rescaling has been widely used in population genetics (McVean and Charlesworth 2000; Tachida 2000; Comeron and Kreitman 2002; Keightley and Eyre-Walker 2007; Kaiser and Charlesworth 2009). In our own experience, rescaling produced reasonable results even when $N$ was set to 10 (see supplemental Table S1).

When computing the more complex model given in Equation 13, it was not feasible to let the simplex algorithm to examine the full domain of $g$ (i.e., $g \in (0, \infty)$), so we restricted the search to $g \in (0.05, 20)$.

**Simulation methods:** To generate a random sample of size $n$ under the constant population size model ($L_0$), we first solved Equation 6 numerically to obtain $f$. We then determined the population frequency of $A$, denoted as $x$, by sampling from $f$. Finally we obtained the number of copies of $A$ in a sample of size $n$ by drawing from a binomial distribution with parameters $n$ and $x$. A similar procedure was used to generate random samples under the more complex model ($L_1$).

**Source of data:** We applied these methods to a large *D. melanogaster* polymorphism dataset (Shapiro et al. 2007). The data, which were kindly provided by D. Turiissini, contained alignments of 468 autosomal loci. A number of Zimbabwe and cosmopolitan lines were sampled. We only used the Zimbabwe lines in our analysis because of severe bottleneck effects in the other population that was sampled (Glinka et al. 2003; Haddrill et al. 2005;
LI and STEPHAN 2006; THORNTON and ANDOLFATTO 2006). Among the Zimbabwe lines, we only used those showing the strongest evidence for reproductive isolation, the “strong Z-lines.” This avoids the possibility that the Z lines with reduced reproductive isolation with the cosmopolitan lines might have been formed by secondary contact with the cosmopolitan lines; these gave an excess of intermediate-frequency variants in the frequency spectrum when these “weak” Z-lines were included in the analysis (J. SHAPIRO and C.-I. WU personal communication).

To annotate the data, we downloaded version 5.11 of the \textit{D. melanogaster} genome data from FlyBase (http://flybase.org/). We then used BLAT (version 34, with default parameters; http://genome-test.cse.ucsc.edu/~kent/exe/linux/) to align the data to the reference genome. We excluded eight loci falling into putatively repetitive regions, and one locus falling into a heterochromatic region. For the other loci, we used their genomic locations to obtain estimates of local recombination rate using a well-established method (SINGH \textit{et al.} 2005) (the data were available online from the PETROV laboratory at http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl). To reduce the effect of linkage, we only analyzed loci locating in highly recombining regions, defined as regions with a local recombination rate strictly higher than 2.3cM/Mb (SINGH \textit{et al.} 2005). To avoid the complication of alternatively spliced genes (IIDA and AKASHI 2000), we only retained loci overlapping one annotated well-defined coding region (i.e., starting with a start codon, ending with a stop codon, and possessing no premature stop codon). We further excluded codons where mutations in the first position could change the status of the third codon position (e.g. from four-fold degenerate to two-fold degenerate). Insertions and deletions were also
removed. For the genes passing all the above filters, we extracted two-fold degenerate codons, and defined preferred/unpreferred codons as in DURET and MOUCHIROUD (1999). A summary of the data is given in Table 1.
RESULTS

Effects of a change in population size on patterns of codon usage evolution: In this section, we assumed that the population size changed instantly from \( N_b \) to \( N_a \) at time zero, and stayed constant thereafter. Allele \( A \) was the preferred allele, but mutation was biased towards the unpreferred allele, \( a \) (i.e. \( \kappa > 1 \)). First, we explored the effects of population expansion by assuming that the population size instantly increased ten-fold at time zero (i.e. \( N_a = 10N_b \); Figure 1). As reported previously (Takano-Shimizu 1999; Charlesworth and Eyre-Walker 2007), for a long period after the expansion (about \( 38N_a \) generations in Figure 1A), the total substitution rate is higher than that before the expansion. This is due to a sharp increase in \( r_{aA} \) (the top curve in Figure 1A). Patterns of polymorphism are also complex (Figure 1B). For example, the frequency of polymorphisms arising from sites fixed for \( a \) (\( P_a \)) first increases, and then gradually decreases to its new equilibrium level. This behavior of \( P_a \) is caused by the following dynamics. Just after the expansion, there are many more sites fixed for \( a \) in the population than at equilibrium under the new, larger population size. With the parameter values used in Figure 1, 68% of the sites are expected to be fixed for \( a \) before the expansion, whereas this number decreases to 12% under the new population size. Thus, immediately after the expansion, the chance of occurrence of an \( a \rightarrow A \) mutation is higher than at equilibrium with the new population size. On the other hand, the population expansion changes the scaled selection coefficient from \( \gamma_b = 2N bs = 0.3 \), the value before the change, to \( \gamma_a = 2N as = 3 \), the value after the change. As a result, under the new population size, \( A \rightarrow a \) polymorphic mutations become more deleterious and are more efficiently purged, whereas
$a \rightarrow A$ polymorphic mutations become more advantageous and are more likely to escape stochastic loss and to segregate at an appreciable frequency. These non-equilibrium dynamics lead to the transient increase in $P_a$. However, as time elapses, the number of sites fixed for $a$ gradually decreases, and $P_a$ also slowly decreases to its new equilibrium value.

Another feature of Figure 1B is of interest: for about $5N_a$ generations’ time $P_a \geq P_A$. This observation is striking because (i) $P_A = P_a = 50\%$, which occurs twice in Figure 1B when the two curves intersect, is a pattern expected under an equilibrium population with neutrality (McVean and Charlesworth 1999) (ii) $P_a > P_A$ is expected under an equilibrium model when $a$ is preferred (McVean and Charlesworth 1999), but here $a$ was unpreferred. These results strongly suggest that using predictions made by equilibrium models of codon usage to interpret patterns of polymorphism observed in a non-equilibrium population may lead to very misleading conclusions.

Next, we examined the effects of a reduction in population size. In Figure 2, a ten-fold decrease was assumed (i.e. $10N_a = N_b$). Immediately after the reduction, the total substitution rate, $r$, is much higher than that before (Figure 2A), consistent with previous results (Takano-Shimizu 1999; Charlesworth and Eyre-Walker 2007). The increase in $r$ is due to increases in both $r_{Aa}$ and $r_{aA}$. Patterns of polymorphism also show interesting dynamics. Before reaching their new equilibrium values of 54% and 46%, respectively, the difference between $P_A$ and $P_a$ increases temporarily (Figure 2B), mimicking an increase in selection pressure on synonymous codons, despite the fact that the selection pressure is greatly reduced due to the sharp decrease in population size. This increased difference between $P_A$ and $P_a$ is caused by (i) a transient increase in the number of sites fixed for $A$ (see supplemental Figure...
S2), which allows more \( A \rightarrow a \) polymorphic mutations to occur (ii) a decrease in the total number of \( a \rightarrow A \) mutations arising each generation. Interestingly, the transient increase in the number of sites fixed for \( A \) also results in a brief increase in \( r_{Aa} \) at around \( 2N_a \) generations after the contraction.

In brief, the dynamics of codon usage can be rather complex following a recent change in population size. Patterns of polymorphism may imitate those expected under a weaker (Figure 1B) or stronger (Figure 2B) selection pressure on synonymous codons. Further, these misleading patterns can persist for long periods, as the time required for the population to reach the new equilibrium state is typically very long (e.g. Figures 1 and 2).

**The effects of a recent population size change on methods that assume equilibrium:**

A number of methods have been proposed to quantify the selection pressure on synonymous codons using within-species polymorphism data (Akashi and Schaeffer 1997; Maside et al. 2004; Comeron and Guthrie 2005; Cutter and Charlesworth 2006). These methods assume that the population is at mutation-selection-drift equilibrium. However, given the results described above, knowing the performance of these methods in a non-equilibrium population is important. To this end, we used the method of Maside et al. (2004) as an example due to its simplicity in calculation. This method relies on some well-established diffusion approximations (McVean and Charlesworth 1999). Our simulations suggest that when the population is at equilibrium and there is mutational bias, the method may give slightly biased estimates of \( \gamma \), but this is apparent only when selection is very weak (see supplemental Figure S3).
In Figure 3, the same model and parameters used to obtain Figure 1 were used to generate random samples of size 15 with 10,000 codons at various time points after the population expansion. We used the MASIDE et al. (2004) method to analyze these samples. As shown by Figure 3A, the mean value of the maximum likelihood estimate (mean MLE) of $\gamma$ starts at a value around 0.3, the true value before the expansion, and then goes down to a value of about -0.5; thereafter, the mean MLE increases monotonically towards the new equilibrium value of 3. Over this time period, the mean MLE of $\gamma$ takes the value of zero twice, at $\sim N_a$ and $\sim 16N_a$ generations after the expansion, respectively. These two special time points correspond to the two occasions when $P_A = P_a = 50\%$ in Figure 1B. The power of the method to reject neutrality at a significance level of 5% is shown in Figure 3B. There are two dips in the power curve. Not surprisingly, these two significant reductions in power occur when the mean MLE of $\gamma$ takes the value of zero. One feature of special note in Figure 3 is that, during the time between the two occasions when the mean MLE of $\gamma$ is zero, the MASIDE et al. (2004) method has great power to reject neutral evolution for the datasets that we simulated. However, the $\gamma$ value that the method returns has, on average, a different sign from the true value. In other words, the preferred/unpreferred state estimated by the method is the reverse of the true situation. Further investigation suggests that the negative mean MLE was not caused by our incorporation of a strong mutational bias in the simulation ($\kappa = 3$), because the same pattern was observed when $\kappa$ was set to one (results not shown).

In agreement with the results shown in Figure 2B, there is a brief increase in the mean MLE of $\gamma$ immediately after a reduction in population size (Figure 4), and thereafter the mean MLE decreases steadily towards the new equilibrium value of 0.2. The power of the MASIDE
et al. method to reject neutrality in this case is a monotonically decreasing function of time (see supplemental Figure S4).

We also carried out simulations to examine effects of non-equilibrium evolution on the AKASHI-SCHAEFFER method (1997; see supplemental Figures S8 and S9) and the CUTTER-CHARLESWORTH method (2006; see supplemental Figures S10 and S11). Both methods are affected. Interestingly, the AKASHI-SCHAEFFER method overestimates the intensity of selection after a population size expansion, and falsely infers the sign of $\gamma$ after a population size reduction; these patterns are the reverse of those for the MASIDE et al. method. The CUTTER-CHARLESWORTH method has similar properties to the MASIDE et al. method.

Taken together, the results suggest that recent population size changes can cause methods that assume equilibrium to give misleading results. First, the estimated selection pressure on synonymous codons may be very different from the true value (Figures 3A and 4); second, neutrality may be falsely accepted (Figure 3B).

The performance of the new estimation methods: To model the effects of a change in population size with the new methods described in Materials and Methods, we constructed two statistical models: $L_0$ and $L_1$. $L_0$ assumes that the population is at equilibrium, and has three parameters: $\gamma$, $\theta$, and $\kappa$. $L_1$ assumes that the population is originally at equilibrium with population size $N_b$. Then the population size changes instantly to $N_a$, and stays constant thereafter for $t$ generations, at which point a sample is taken from the population. Hence, $L_1$ has five parameters: $\gamma_b (= 2N_b\delta)$, $\theta_b (= 4N_b\mu)$, $\kappa$, $g (= N_a / N_b)$, and $\tau (= t / N_a)$.

In contrast to previous methods, which do not estimate mutational parameters (AKASHI
and Schaeffer 1997; Maside et al. 2004; Comeron and Guthrie 2005; Cutter and Charlesworth 2006), L₀ simultaneously estimates mutational (θ and κ) and selection (γ) parameters. Simulation results show that L₀ has high accuracy in estimating all the parameters (Table 2). However, similarly to the Maside et al. method, L₀ is susceptible to violation of the equilibrium assumption (see supplemental Figures S5 and S6). In this case, the behavior of L₀ is rather similar to that of the Maside et al. method: following a recent population expansion the mean MLE of γ first decreases and then slowly increases to the new equilibrium value. The difference is that the mean MLE of γ returned by L₀ takes a minimum value of about 0.1, and never becomes negative. Estimates of κ and θ produced by L₀ are also likely to be biased by recent demographic changes (see supplemental Figures S5 and S6).

The L₁ method has two more parameters, τ and g, and requires the iteration of Equation 5. Hence, L₁ is the most computationally intense method considered here. Again we did simulations to examine the properties of L₁ (Table 3). The accuracy of L₁ in estimating the parameters is generally quite high. Importantly, the method seems to have good power to reject a model with constant population size and to detect selection (Table 3). However, when only a relatively small amount of data is available, the accuracy and power of the L₁ method are reduced (see the second set of results in Table 3). In this case, the model with a recent population size reduction seems to be particularly problematic: in the simulations, the simplex algorithm we used often failed to converge, and returned unrealistic results. Interestingly, reducing the amount of data does not reduce the power of the method to detect selection if the population size is declining, whereas limited amount of data does reduce the power to detect selection if the population size is increasing. This difference is probably
caused by the fact that immediately after a reduction in population size, patterns of polymorphism can be very selection-like (i.e. \( P_A \) and \( P_a \) are very different; e.g. Figure 2B), whereas after a population expansion patterns of polymorphism can be more neutral-like (i.e. \( P_A \) and \( P_a \) are close to 50%; e.g. Figure 1B).

Overall, the \( L_1 \) method seems to have substantial power to disentangle selection from demography. This is a useful property because it has long been known that selection and demography, either acting separately or jointly, can leave very similar traces on polymorphism patterns, and distinguishing them has been a subject of active research (Simonsen et al. 1995; Nielsen et al. 2005; Williamson et al. 2005; Keightley and Eyre-Walker 2007; Zeng et al. 2007).

Applying the new methods to a \( D. \) melanogaster dataset: We used our new methods, \( L_0 \) and \( L_1 \), to analyze a large \( D. \) melanogaster polymorphism dataset (Shapiro et al. 2007), summarized in Table 1 (see also Materials and Methods).

Initial analysis: The results of the analyses are given in Table 4. Interestingly, the simple model, \( L_0 \), seems to be sufficient to explain the data — there is essentially no increase in the log-likelihood (\( \ln L \)) when the data was analyzed using \( L_1 \). Further support for this conclusion comes from the fact that, under \( L_1 \), the estimate of the time after expansion, \( \tau \), is very large, by which time the population has essentially reached a new equilibrium, and \( L_1 \) reduces to \( L_0 \). For this reason, we were unable to obtain a 95% confidence interval under \( L_1 \). Furthermore, consistent with many previous reports (Petrov and Hartl 1999; McVean and Vieira 2001; Nielsen et al. 2007; Keightley et al. 2009), the mutational process is not
symmetric, but is biased towards preferred-to-unpreferred changes ($L_0(\kappa = 1)$ versus $L_0$, $\chi^2 = 4.6$, $d.f. = 1$, $p = 0.03$). In contrast, comparing $L_0$ with the reduced model when the selection coefficient ($s$) was fixed at zero (i.e. $L_0(s = 0)$ versus $L_0$), we found strong evidence for recent selection on synonymous codons ($\chi^2 = 19.7$, $d.f. = 1$, $p = 9.1 \times 10^{-6}$). Using the estimated parameters, the $L_0$ model can accurately predict the observed data (see supplemental Figure S7). In short, the above results suggest that a model with constant population size and selection acting on synonymous codons is a sufficient explanation of the data.

A more detailed analysis: Although the $L_0$ model seems to provide a rather good fit to the data, it is highly simplified in that it assumes a single selection coefficient and a single mutational bias parameter for all sites. It has been shown by other researchers that different genes can experience different levels of selection pressure, depending on factors such as gene expression level and length of coding region (Comeron and Kreitman 2002; Maside et al. 2004; Comeron and Guthrie 2005; Cutter and Charlesworth 2006), and that the mutational bias parameter can be different for different nucleotides (McVean and Vieira 1999; Petrov and Hartl 1999; Keightley et al. 2009). We therefore introduced additional parameters into the $L_0$ model.

First, some of the preferred codons being analyzed differ from their corresponding unpreferred codons by a G to A change at the third codon position, while the rest differ by C to T. We thus extended the model so that each class of codons had its own set of parameters. A maximum likelihood analysis suggests that this model fits the data much better than the original model ($\chi^2 = 287.0$, $d.f. = 3$, $p < 10^{-20}$). The MLEs are $\gamma_{GA} = 0.85$, $\gamma_{CT} = 1.13$, $\hat{\theta}_{GA} = 0.011$, $\hat{\theta}_{CT} = 0.007$, $\kappa_{GA} = 0.95$, and $\kappa_{CT} = 2.26$ ($\ln L = -11484.3$). However, a model
with a common selection coefficient for the two classes of codons fits the data almost equally well ($\chi^2 = 0.34, d.f. = 1, p = 0.56$); the MLEs and the log-likelihood under this model are: $\hat{\gamma} = 1.03, \hat{\theta}_{GA} = 0.010, \hat{\theta}_{CT} = 0.008, \hat{\kappa}_{GA} = 1.13, \hat{\kappa}_{CT} = 2.04$, and $\ln L = -11484.5$. Thus, we assumed that there was no difference in selection coefficient between G/A or C/T ending codons in the following more elaborate analysis.

Another major determinant of codon usage is the level of gene expression (Duret and Mouchiroud 1999; Hey and Kliman 2002). To model this complication, we used data from the UniGene database (http://www.ncbi.nlm.nih.gov/unigene), employing the total number of ESTs (expressed sequence tags) for a gene as a rough measure of its level of expression (the data were kindly provided by B. Vicoso). We divided the genes into two equal-size groups: low-expression and high-expression genes. We allowed each group of genes to have their own scaled selection coefficient, $\gamma_{\text{low}}$ and $\gamma_{\text{high}}$, respectively. By fitting the model to the data, we obtained the following MLEs: $\hat{\gamma}_{\text{low}} = 0.82, \hat{\gamma}_{\text{high}} = 1.25, \hat{\theta}_{GA} = 0.010, \hat{\theta}_{CT} = 0.008, \hat{\kappa}_{GA} = 1.14, \hat{\kappa}_{CT} = 2.03$, and $\ln L = -11079.3$. This new model further improves the fit to the data compared to $L_0(d|\gamma, \theta_{GA}, \theta_{CT}, \kappa_{GA}, \kappa_{CT})$ ($\chi^2 = 810.3, d.f. = 1, p < 10^{-20}$). In accordance with the previous reports, our results suggest that highly expressed genes are under stronger selection pressure than genes with low expression levels.

Finally, we considered the effects of coding region length. We divided low- and high-expression genes, respectively, into two equal-size classes: genes with short and long coding regions. We allowed each group of genes to have its own selection coefficient. The MLEs are: $\hat{\gamma}_{\text{low-short}} = 0.91, \hat{\gamma}_{\text{low-long}} = 0.76, \hat{\gamma}_{\text{high-short}} = 1.36, \hat{\gamma}_{\text{high-long}} = 1.17, \hat{\theta}_{GA} = 0.010, \hat{\theta}_{CT} = 0.008, \hat{\kappa}_{GA} = 1.14, \hat{\kappa}_{CT} = 2.04$, and $\ln L = -11068.6$. The improvement in
the fit to the data, compared to the model given in the previous paragraph, is also highly
significant ($\chi^2 = 21.3$, d.f. = 2, $p = 2.3 \times 10^{-5}$). These results confirm the finding that genes with
shorter coding regions tend to be under stronger selection pressure than those with longer
coding regions (COMERON et al. 1999; DURET and MOUCHIROUD 1999; COMERON and
GUTHRIE 2005).
DISCUSSION

The importance of considering non-equilibrium factors in the study of codon usage bias: Population-genetic models of codon usage bias (LI 1987; BULMER 1991; McVEAN and CHARLESWORTH 1999) have mostly assumed that the population is at mutation-selection-drift equilibrium. As evidence for non-equilibrium situations in Drosophila species accumulates (TAKANO-SHIMIZU 1999, 2001; KERN and BEGUN 2005; LI and STEPHAN 2006; THORNTON and ANDOLFATTO 2006; STEPHAN and LI 2007), the study of non-equilibrium models has attracted more attention, and some unexpected results have emerged (EYRE-WALKER 1997; TAKANO-SHIMIZU 1999, 2001; COMERON and KREITMAN 2002; COMERON and GUTHRIE 2005; CHARLESWORTH and EYRE-WALKER 2007).

Our matrix model allows systematic studies of the patterns of polymorphism and substitution following a recent change in population size. We were able to confirm and extend the findings reported in the previous studies. Most importantly, we have carried out the first survey of the performance of statistical methods that assume equilibrium in a situation where such an assumption is violated. We can draw two conclusions. First, patterns of polymorphism can be very misleading after a recent change in population size. For example, they can become very neutral-like after a recent population expansion (Figure 1B). In contrast, after a recent population size reduction, patterns of polymorphism are likely to suggest a level of selection pressure higher than the true value (Figure 2B). Second, using statistical methods that assume equilibrium in a non-equilibrium population can result in erroneous inferences of the strength of selection (Figures 3 and 4; see also supplemental Figures S5, S6, S8 – S11).
These findings indicate that it is important to consider non-equilibrium factors in the study of codon usage bias.

**Effects of weak selection on synonymous variants:** Contrary to a previous analysis of the same *D. melanogaster* dataset (Keightley and Eyre-Walker 2007), we did not find any evidence for a recent population expansion in the Zimbabwe sample. The main difference between the studies is that synonymous sites were previously assumed to be neutral. However, as shown here, there is evidence for selection on these sites. According to standard population genetics theory, such a selection pressure can result in an excess of low-frequency variants (Ewens 2004, Chap. 5). It is likely that this excess of low-frequency variants was treated by the method of Keightley and Eyre-Walker as the signature of a recent population size expansion. Indeed, when these authors fitted the data using a model with selection on synonymous sites, the support for the expansion event disappeared (see p. 2260 in Keightley and Eyre-Walker 2007). Moreover, recent studies have shown that non-coding regions may be under selection as well (e.g. Andolfatto 2005). These findings suggest that treating certain types of markers as neutral standards can be a dubious procedure, and deserves careful investigation of its validity, whose violation may result in unreliable inferences.

**Selection on synonymous codons in D. melanogaster:** By analyzing a large polymorphism dataset (Shapiro et al. 2007), we found evidence for natural selection on synonymous codons in *D. melanogaster* (Table 4). Furthermore, we show that genes with
shorter coding sequences and/or higher levels of expression are under stronger selection. These results are consistent with previous studies, which suggest that patterns of codon usage bias in *D. melanogaster* are compatible with the operation of selection (KLIMAN 1999; HEY and KLIMAN 2002; CARLINI and STEPHAN 2003; QIN et al. 2004; COMERON and GUTHRIE 2005; SINGH et al. 2007).

However, a number of studies have also found that selection on synonymous sites may have been substantially reduced in the *D. melanogaster* lineage (AKASHI 1995, 1996; MCVEAN and VIEIRA 1999, 2001; NIELSEN et al. 2007). A common feature of these studies is that they used between-species divergence data, and drew conclusions based on the observation that *D. melanogaster* synonymous sites have fixed significantly more unpreferred alleles than preferred variants (the rates of fixation should be equal at equilibrium). Divergence data can tell us what has happened along the entire *D. melanogaster* lineage, but not when these events happened. To reconcile these findings with ours, we hypothesize that there may have been a reduction in population size in the *D. melanogaster* lineage a long time ago, which resulted in the rapid fixation of many unpreferred variants (e.g. Figure 2A). After this ancient reduction, the population size may have stayed relatively constant, and extant patterns of polymorphism may have come relatively close to the new equilibrium. Because our method does not use divergence data, it has little power to detect such an ancient event, so that the equilibrium model provides a good fit.

The difficulty is that, even if this hypothesis is correct, it is hard to estimate when the reduction in population size occurred, and our method may thus overestimate the current intensity of selection, depending on the time of the contraction event (see supplemental
Figure S6). In addition, the potential influences of biased gene conversion (Marais et al. 2001; Kliman and Hey 2003; Maraiss 2003; Maraiss et al. 2003), changes in the recombinational landscape (Takano-Shimizu 1999), and recent changes in the mutational processes (Takano-Shimizu 2001; Kern and Begun 2005), as well as the effects of more complex demographic models (e.g. population structure and population bottlenecks), are not considered by our method. Therefore, although there is evidence for selection on synonymous sites, this result should be treated with caution.

Advantages and caveats: The matrix model and the methods developed in this study have a number of advantages. First, the model is highly flexible. Although we focus on the effects of a sudden change in population size here, it is straightforward to modify the model so that it can be used to study effects of more complex demographic models (e.g. population bottlenecks, exponential growth), changes in selection coefficients, and changes in mutational parameters. Second, our methods allow the simultaneous estimation of selection and mutational parameters, in contrast to previous methods that rely only on polymorphic sites (Maside et al. 2004; Cutter and Charlesworth 2006). Third, the statistical methods we have proposed do not require the use of an outgroup sequence to infer ancestral versus derived states, a process which is error-prone and can lead to unreliable inferences.

There are also some caveats. First, as most previous studies, we assume that sites are independent. This assumption is obviously unrealistic. Hill-Robertson effects among linked sites might make our methods unreliable (McVean and Charlesworth 2000; Comeron et al. 2008). Fortunately, recent simulation studies have suggested that methods
assuming free recombination work surprisingly well in practice with relatively free recombination (e.g. with a local recombination rate higher than 2cM/Mb, K. ZENG unpublished results; see also WILLIAMSON et al. 2005; and BOYKO et al. 2008). By using data from highly recombining regions, our results should be robust to the violation of the free-recombination assumption. The second disadvantage of the matrix model is that it is computationally burdensome. In particular, the $L_1$ model, which requires iterating Equation 5, can be very slow. This is a common problem confronting all studies using matrix models (e.g. KEIGHTLEY and EYRE-WALKER 2007). Despite these caveats, matrix models are flexible and easy to formulate, and their usefulness should not be underestimated.

Program availability: A computer program implementing the methods described in this paper is available from the corresponding author.
ACKNOWLEDGEMENTS

We thank Andrea Betancourt, Penny Haddrill, and Peter Keightley for helpful discussions, and David Turissini for providing the Drosophila dataset. We also thank Rasmus Nielsen, John Wakeley, and the anonymous reviewers for their insightful comments. We are very grateful for the Sun Yat-Sen University Computer Centre for providing computer time, and Hailiang Liu and Haijun Wen for technical support. This work has also made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk). K.Z. was supported by a Royal Society International Incoming Fellowship. K.Z. also acknowledges support from a National Basic Research Program of China grant (2007CB815701), and a National Natural Science Foundation of China grant (30730008) to Professor Suhua Shi.
LITERATURE CITED


AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067-1076.


COMERON, J. M., M. KREITMAN and M. AGUADE, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151: 239-249.


QIN, H., W. B. WU, J. M. COMERON, M. KREITMAN and W. H. LI, 2004 Intragenic spatial


FIGURE LEGENDS

Figure 1. Patterns of substitution and polymorphism after a recent population size expansion. We assumed that a diploid population of size $N_b$ was originally at equilibrium. At time zero the population size increased instantly ten-fold to $N_a$, and stayed constant thereafter. Allele $A$ was assumed to be the preferred allele. The parameters used to generate this figure were: $\gamma_b = 2N_b\delta = 0.3$, $\theta_b = 4N_b\mu = 0.002$, $\kappa = 3$. Time was measured in units of $N_a$ generations. (A) Patterns of substitution. We calculated three quantities: $r_{aA}$ (the rate at which sites fixed for allele $a$ were replaced by sites fixed for allele $A$; see Equation 10), $r_{AA}$ (the rate at which sites fixed for $A$ were replaced by sites fixed for $a$), and $r = r_{aA} + r_{AA}$, the total substitution rate. The figure shows the ratios of the values of these quantities after the change in population size to their equilibrium values before the change. (B) Patterns of polymorphism. $P_A$ (or $P_a$) is the proportion of polymorphic sites originating from sites previously fixed for $A$ (or $a$) (see Equation 9).

Figure 2. Patterns of substitution and polymorphism after a recent population size shrinkage. We assumed that a diploid population of size $N_b$ was originally at equilibrium. At time zero the population size decreased instantly ten-fold to $N_a$, and stayed constant thereafter. Allele $A$ was assumed to be the preferred allele. The parameters used to generate this figure were: $\gamma_b = 2$, $\theta_b = 0.02$, $\kappa = 3$. Time was measured in units of $N_a$ generations. (A) Pattern of substitution. (B) Pattern of polymorphism. Definitions of the quantities, $r_{aA}$, $r_{AA}$, $r$, $P_A$, and $P_a$, are the same as those given in Figure 1.
**Figure 3. Effects of a recent population size expansion on the MASIDE et al. (2004) method.** The model and parameters used to generate this figure were the same as those used to generate Figure 1 (i.e., a ten-fold increase in population size, $\gamma_b = 0.3$, $\theta_b = 0.002$, and $\kappa = 3$). At each time point after the expansion in population size, we randomly generated 500 samples. Each sample was composed of 15 sequences of 10,000 codons. We then used the MASIDE et al. (2004) method to analyze these samples, and obtained a distribution of estimated $\gamma$. (A) The black curve shows the mean value of the estimates of $\gamma$, and the grey sleeve indicates where 95% of the probability mass of the distribution lies. (B) The power of the MASIDE et al. method to reject neutrality at a significance level of 5%. Note that the time scale is different in the two plots.

**Figure 4. Effects of a recent population size reduction on the MASIDE et al. (2004) method.** The model and parameters used to generate this figure were the same as those used to generate Figure 2 (i.e. a ten-fold decrease in population size, $\gamma_b = 2$, $\theta_b = 0.02$, $\kappa = 3$). At each time point after the reduction in population size, we randomly generated 500 samples. Each sample was composed of 15 sequences of 10,000 codons. We then used the MASIDE et al. method to analyze these samples, and obtained a distribution of estimated $\gamma$. The black curve shows the mean of the estimates of $\gamma$, and the grey sleeve indicates where 95% of the probability mass of the distribution lies.
### Table 1. Summary of the *D. melanogaster* polymorphism data.

<table>
<thead>
<tr>
<th>$G^a$</th>
<th>$\bar{n}^b$</th>
<th>$L^c$</th>
<th>$S^d$</th>
<th>$\theta_w^e$</th>
<th>$\theta_\pi^f$</th>
<th>Tajima’s $D^g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>182</td>
<td>5.5</td>
<td>14,807</td>
<td>335</td>
<td>0.0107</td>
<td>0.0108</td>
<td>-0.0076</td>
</tr>
</tbody>
</table>

*a* Total number of loci.

*b* Mean sample size.

*c* Total number of two-fold degenerate sites.

*d* Total number of polymorphic two-fold degenerate sites.

*e* Mean value of Watterson’s estimator of (per site) population mutation rate (*Watterson* 1975); in the calculation, only two-fold degenerate sites were used.

*f* Mean nucleotide diversity; in the calculation, only two-fold degenerate sites were used.

*g* Mean Tajima’s $D$ statistic (*Tajima* 1989); in the calculation, only two-fold degenerate sites were used.
Table 2. Performance of the $L_0$ method in an equilibrium population.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\theta$</th>
<th>$\kappa$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size = 15, number of codons per sequence = 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input$^a$</td>
<td>0.01</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mean MLE$^b$</td>
<td>0.01</td>
<td>1.02</td>
<td>-0.0009</td>
</tr>
<tr>
<td>Percentiles$^c$</td>
<td>[0.008, 0.012]</td>
<td>[0.71, 1.45]</td>
<td>[-0.3772, 0.3488]</td>
</tr>
<tr>
<td>Input$^a$</td>
<td>0.01</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Mean MLE$^b$</td>
<td>0.01</td>
<td>3.01</td>
<td>1.02</td>
</tr>
<tr>
<td>Percentiles$^c$</td>
<td>[0.008, 0.012]</td>
<td>[2.32, 4.10]</td>
<td>[0.72, 1.34]</td>
</tr>
</tbody>
</table>

Sample size = 25, number of codons per sequence = 500

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\theta$</th>
<th>$\kappa$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input$^a$</td>
<td>0.01</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Mean MLE$^b$</td>
<td>0.01</td>
<td>3.53</td>
<td>1.98</td>
</tr>
<tr>
<td>Percentiles$^c$</td>
<td>[0.003, 0.019]</td>
<td>[1.01, 10.62]</td>
<td>[0.79, 3.36]</td>
</tr>
</tbody>
</table>

For each combination of parameter values, we used the matrix model to generate 200 samples. These samples were analyzed using $L_0$ (see Equation 12).

$^a$ The parameter values used to generate random samples.

$^b$ Mean value of the maximum likelihood estimates (MLE).

$^c$ 2.5% and 97.5% percentiles of the distributions of the MLEs.
Table 3. Performance of the $L_1$ method in a non-equilibrium population.

<table>
<thead>
<tr>
<th></th>
<th>$\theta_b$</th>
<th>$\kappa$</th>
<th>$\gamma_b$</th>
<th>$g$</th>
<th>$\tau$</th>
<th>Power1$^d$</th>
<th>Power2$^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input$^a$</strong></td>
<td>0.004</td>
<td>3</td>
<td>0.4</td>
<td>5</td>
<td>1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Mean MLE$^b$</strong></td>
<td>0.0042</td>
<td>3.06</td>
<td>0.41</td>
<td>5.41</td>
<td>0.94</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Percentiles$^c$</strong></td>
<td>[0.0023, 0.0077]</td>
<td>[2.61, 3.65]</td>
<td>[0.26, 0.59]</td>
<td>[3.10, 8.35]</td>
<td>[0.41, 1.32]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Input$^a$</strong></td>
<td>0.02</td>
<td>3</td>
<td>2</td>
<td>0.2</td>
<td>2</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Mean MLE$^b$</strong></td>
<td>0.022</td>
<td>2.69</td>
<td>1.86</td>
<td>0.23</td>
<td>1.86</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Percentiles$^c$</strong></td>
<td>[0.015, 0.043]</td>
<td>[1.62, 4.57]</td>
<td>[1.30, 2.48]</td>
<td>[0.12, 0.32]</td>
<td>[0.68, 3.73]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample size = 25, number of codons per sequence = 25,000

Sample size = 6, number of codons per sequence = 15,000

<table>
<thead>
<tr>
<th></th>
<th>$\theta_b$</th>
<th>$\kappa$</th>
<th>$\gamma_b$</th>
<th>$g$</th>
<th>$\tau$</th>
<th>Power1$^d$</th>
<th>Power2$^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input$^a$</strong></td>
<td>0.004</td>
<td>3</td>
<td>0.4</td>
<td>5</td>
<td>1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Mean MLE$^b$</strong></td>
<td>0.0045</td>
<td>3.42</td>
<td>0.48</td>
<td>5.9</td>
<td>0.90</td>
<td>98%</td>
<td>57%</td>
</tr>
<tr>
<td><strong>Percentiles$^c$</strong></td>
<td>[0.0018, 0.0097]</td>
<td>[2.14, 6.23]</td>
<td>[0.06, 1.09]</td>
<td>[3.0, 9.2]</td>
<td>[0.06, 1.87]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input(^a)</td>
<td>0.02</td>
<td>3</td>
<td>2</td>
<td>0.2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean MLE(^b)</td>
<td>0.027</td>
<td>3.05</td>
<td>1.9</td>
<td>0.21</td>
<td>2.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentiles(^c)</td>
<td>[0.008, 0.095]</td>
<td>[1.41, 9.06]</td>
<td>[1.24, 3.31]</td>
<td>[0.08, 0.44]</td>
<td>[0.73, 5.75]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the simulations, we assumed that the population was originally at equilibrium with population size \(N_b\). Then the population size changed instantly to \(N_a\), and stayed constant thereafter for \(t\) generations, at which point random samples were taken. The parameters were: \(\gamma_b (= 2N_b s)\), \(\theta_b (= 4N_b u)\), \(\kappa\), \(g (= N_a / N_b)\), and \(\tau (= t / N_a)\). For each combination of parameter values, 100 replicate simulations were run.

\(^a\) The parameter values used to generate random samples.

\(^b\) Mean value of the maximum likelihood estimates (MLE).

\(^c\) 2.5% and 97.5% percentiles of the distributions of the MLEs.

\(^d\) The power to reject a model with constant population size at a significance level of 5%.

\(^e\) The power to reject neutral evolution at a significance level of 5%.
Table 4. Analyzing the two-fold degenerate codons using $L_0$ and $L_1$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimates of parameters</th>
<th>log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_b$</td>
<td>$\kappa$</td>
</tr>
<tr>
<td>$L_1$ MLE</td>
<td>0.0040</td>
<td>1.67</td>
</tr>
<tr>
<td>$L_0$ MLE</td>
<td>0.0083</td>
<td>1.63</td>
</tr>
<tr>
<td>95% CI$^a$</td>
<td>[0.0061, 0.0109]</td>
<td>[1.03, 2.54]</td>
</tr>
<tr>
<td>$L_0(s = 0)^b$</td>
<td>MLE</td>
<td>0.0144</td>
</tr>
<tr>
<td>$L_0(\kappa = 1)^c$</td>
<td>MLE</td>
<td>0.0110</td>
</tr>
</tbody>
</table>

$^a$ 95% confidence interval obtained by 250 bootstrap replicates.

$^b$ The $L_0$ model with $s$, the selection coefficient, fixed at zero.

$^c$ The $L_0$ model with $\kappa$, the mutational bias parameter, fixed at unity.
Figure 1A.

Figure 1B.
Figure 2A.

![Figure 2A](image)

Figure 2B.

![Figure 2B](image)
Figure 3A.

![Graph showing the function E(γ) over time after expansion in units of N_a generations.]

Figure 3B.

![Graph showing power over time after expansion in units of N_a generations.]
Figure 4.

![Graph showing $E(\gamma)$ over time after shrinkage in units of $N_a$ generations.](image)