

# Estimating Selection on Nonsynonymous Mutations

Laurence Loewe,<sup>1</sup> Brian Charlesworth, Carolina Bartolomé<sup>2</sup> and Véronique Noël

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

Manuscript received June 23, 2005  
Accepted for publication November 5, 2005

## ABSTRACT

The distribution of mutational effects on fitness is of fundamental importance for many aspects of evolution. We develop two methods for characterizing the fitness effects of deleterious, nonsynonymous mutations, using polymorphism data from two related species. These methods also provide estimates of the proportion of amino acid substitutions that are selectively favorable, when combined with data on between-species sequence divergence. The methods are applicable to species with different effective population sizes, but that share the same distribution of mutational effects. The first, simpler, method assumes that diversity for all nonneutral mutations is given by the value under mutation-selection balance, while the second method allows for stronger effects of genetic drift and yields estimates of the parameters of the probability distribution of mutational effects. We apply these methods to data on populations of *Drosophila miranda* and *D. pseudoobscura* and find evidence for the presence of deleterious nonsynonymous mutations, mostly with small heterozygous selection coefficients (a mean of the order of  $10^{-5}$  for segregating variants). A leptokurtic gamma distribution of mutational effects with a shape parameter between 0.1 and 1 can explain observed diversities, in the absence of a separate class of completely neutral nonsynonymous mutations. We also describe a simple approximate method for estimating the harmonic mean selection coefficient from diversity data on a single species.

**S**URVEYS of DNA sequence polymorphisms in many species have revealed substantial variation in the amino acid sequences of proteins, although the nonsynonymous nucleotide site diversity is usually much less than that for silent variants (LI 1997). This is consistent with the action of purifying selection on protein sequences, removing deleterious amino acid mutations from the population while neutral or nearly neutral silent variants can persist (KIMURA 1983; LI 1997). It is clearly important to characterize the distribution of fitness effects of new nonsynonymous mutations. This distribution is relevant to a broad range of problems in evolutionary genetics, and a variety of methods have been used to characterize it, including direct estimates from mutation-accumulation experiments and indirect estimates from comparisons of DNA sequences among related species (KEIGHTLEY and EYRE-WALKER 1999). The extent, nature, and magnitude of selection on amino acid variants are also relevant to understanding the relation between human disease and genetic variation (SUNYAEV *et al.* 2001; WRIGHT *et al.* 2003).

Several methods have been used to detect purifying selection from data on variability in natural populations

and to estimate the parameters describing such selection. An important method was introduced by SAWYER and HARTL (1992), based on the McDonald-Kreitman test (MCDONALD and KREITMAN 1991). It compares the ratio of the number of within-species nonsynonymous polymorphisms to the number of synonymous polymorphisms and the corresponding ratio for fixed differences between a pair of related species. This ratio of ratios is called the “neutrality index” (RAND and KANN 1996). It is expected to be greater than one if there is predominantly purifying selection against nonsynonymous variants, since selection is less effective in reducing the level of polymorphism for deleterious variants than in preventing their fixation (KIMURA 1983). This approach can be incorporated into maximum-likelihood or Bayesian methods for estimating  $N_e s$ , where  $N_e$  is the effective population size and  $s$  is the selection coefficient against nonsynonymous mutations. The selective effects of mutations are usually assumed to be codominant (in the case of nuclear genes), and constant across sites within a gene (*e.g.*, NACHMAN 1998; BUSTAMANTE *et al.* 2002). Recent extensions allow for a distribution of selection coefficients at different sites (PIGANEAU and EYRE-WALKER 2003) or varying degrees of dominance (WILLIAMSON *et al.* 2004).

Overall, this method has been more successful in detecting and estimating purifying selection on nonsynonymous variants in mitochondrial genomes than in the nuclear genome (NACHMAN 1998; RAND and KANN 1998; WEINREICH and RAND 2000), with a few exceptions such

<sup>1</sup>Corresponding author: Institute of Evolutionary Biology, School of Biological Sciences, Ashworth Laboratories, University of Edinburgh, King's Bldgs., Edinburgh EH9 3JT, United Kingdom.  
E-mail: laurence.loewe@evolutionary-research.net

<sup>2</sup>Present address: Unidade de Xenética Evolutiva, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.

as *Arabidopsis thaliana* (BUSTAMANTE *et al.* 2002), *Drosophila miranda* (BARTOLOMÉ *et al.* 2005), and humans (BUSTAMANTE *et al.* 2005). The fixed selection coefficient model was fitted by NACHMAN (1998) to 17 animal mitochondrial DNA data sets with neutrality index values  $>1$  and gave  $N_e s$  estimates predominantly between 1 and 3; BUSTAMANTE *et al.* (2002) estimated  $N_e s$  on a gene-by-gene basis and obtained a mean value of  $\sim 1$  for 12 nuclear genes in *A. thaliana* (in both cases,  $N_e$  is the haploid effective population size, as is appropriate for the ploidy and breeding systems in these cases). A modification of this approach, allowing a normal distribution of  $N_e s$  values across different genes, was applied to data on *D. melanogaster*, indicating a mean  $N_e s$  of  $\sim 3.5$  (SAWYER *et al.* 2003). The fits of a gamma distribution of selection coefficients across individual nonsynonymous sites to animal mitochondrial data sets (PIGANEAU and EYRE-WALKER 2003) gave much larger but noisier estimates of mean  $N_e s$ .

An alternative approach is to fit the observed frequency spectra of nonsynonymous and silent/synonymous variant frequencies to the distributions expected under mutation-selection-drift equilibrium (SAWYER 1994; SAWYER *et al.* 1987). Variants of this approach have been developed by HARTL *et al.* (1994), AKASHI (1999), FAY *et al.* (2001), BUSTAMANTE *et al.* (2003), and WILLIAMSON *et al.* (2004, 2005). Either a fixed given  $N_e s$  value or a distribution of  $N_e s$  values across loci is assumed.

These studies have yielded large differences among estimates of the scaled selection parameter  $N_e s$  (or its arithmetic mean) for nonsynonymous sites under purifying selection, ranging from values of the order of 1 to several hundred, depending on the methods and species used. Varying conclusions about the proportion of sites subject to positive, as opposed to purifying selection, have also been reached; for example, compare SAWYER *et al.* (2003) with BIERNE and EYRE-WALKER (2004), who estimated that 94 and 25% of nonsynonymous mutations distinguishing *D. simulans* and *D. melanogaster* were fixed by positive selection, respectively.

Methods that fit details of frequency spectra to equilibrium models are clearly highly vulnerable to departures from equilibrium, and there is increasing evidence that many of the model systems used for the study of natural variation, as well as human populations, are subject to such effects (ANDOLFATTO and PRZEWSKI 2000; WILLIAMSON *et al.* 2005). Incorporating even the simplest model of demographic change into selection models is computationally extremely demanding (WILLIAMSON *et al.* 2005). Methods that ignore the details of the distribution of variant frequencies may thus be preferable to potentially more powerful methods that exploit all the features of the data. Another problem is that many of the methods outlined above assume that amino acid mutations in a given gene are unidirectionally from wild type to selectively deleterious

or vice versa. However, unless the magnitude of  $N_e s$  for all mutations is much greater than one, there will be a flux of amino acid substitutions over evolutionary time, such that some fraction of sites will be fixed for selectively deleterious alleles and can therefore back mutate to create fitter variants, as in models of codon usage bias (LI 1987; BULMER 1991; MCVEAN and CHARLESWORTH 1999). Only the model of PIGANEAU and EYRE-WALKER (2003) has explicitly incorporated this possibility.

In this article, we develop a model of nonsynonymous site variation and evolution that includes reversible mutation, as in the standard models of codon usage evolution. We use this to estimate the strength of selection on amino acid variants, by exploiting the difference in the responses of nonsynonymous and synonymous/silent variants to differences in effective population sizes between related species. The basic idea is that variants subject to sufficiently strong purifying selection will not increase much in abundance as effective population size increases, whereas neutral or nearly neutral diversity is expected to increase in proportion to  $N_e$ . The extent to which nonsynonymous diversity differs between species with different synonymous diversity values should thus shed light on the prevalence and strength of purifying selection. We also show how to provide useful bounds on selection parameters when data on only one species are available.

## MATERIALS AND METHODS

**Source of data:** We used published DNA sequence information on X-linked and autosomal genes for *D. miranda* (Yi *et al.* 2003; BARTOLOMÉ *et al.* 2005), removing genes for which there was either evidence for departure from neutrality (*Annx*, *swallow*) from the HKA test or no sequence data from *D. affinis*, the outgroup species used to estimate interspecific divergence from *D. miranda* and *D. pseudoobscura* (BARTOLOMÉ *et al.* 2005). Polymorphism data on *D. pseudoobscura* were compiled from population surveys; the gene *exu2* was not used, since it showed evidence for selection on the basis of haplotype structure (Yi *et al.* 2003) and the HKA test (this study; data not shown). These data were obtained from GenBank accessions, and sequences were aligned using the program SeAl (<http://evolve.zoo.ox.ac.uk/>). Details of the genes used and the relevant references are provided in supplemental material at <http://www.genetics.org/supplemental/>.

In total, 17 *D. miranda* genes (13,309 nonsynonymous sites and 9077 silent sites, 51% in introns) and 14 *D. pseudoobscura* genes (10,828 nonsynonymous sites and 10,245 silent sites, 65% in introns) were used. Sample sizes were 11 or 12 alleles per gene for *D. miranda* and 7–139 per gene for *D. pseudoobscura*. No adjustments for different effective population sizes for X-linked *vs.* autosomal genes were made, as mean diversities are similar for these two categories in both species, consistent with the action of sexual selection on males (see supplementary material at <http://www.genetics.org/supplemental/> and Yi *et al.* 2003; BARTOLOMÉ *et al.* 2005). Polymorphism and divergence estimates were calculated using DnaSP (ROZAS *et al.* 2003). The estimates of divergence

from *D. affinis* for the *D. miranda* loci were obtained from BARTOLOMÉ *et al.*'s (2005) Table 3. Unfortunately, only three loci are in common between the two data sets, so that we have to treat the two sets of genes as representing more or less independent random samples from the genomes of the two species.

**Computational methods:** The case of "arbitrary purifying selection" described in THEORETICAL FRAMEWORK (Equations 5) requires integration of the formulas for the nonsynonymous nucleotide site diversities  $\pi_A$  and rates of substitution  $K_A$ , over an assumed distribution of selection coefficients,  $\phi(s)$ . The relevant expressions involve integrals representing the sojourn times of codominant autosomal mutations, given the heterozygous selection coefficient  $s$ , the breeding adult population size  $N$ , the effective population size  $N_e$ , and the mutation rate per nucleotide site per generation  $u$ . These were obtained from the known solutions to the relevant diffusion equations (KIMURA and OHTA 1969b; McVEAN and CHARLESWORTH 1999). To predict  $\pi_S$ , we used similar equations as for  $\pi_A$ , but setting  $s = 0$ . All computations were implemented in the statistical programming language "R" (version 1.9) (IHAKA and GENTLEMAN 1996; R-PROJECT 2005).

Most mutations generated by a given  $\phi(s)$  distribution have effects that can be handled by the diffusion methods employed here. However, for very strongly or weakly selected mutations, the formulas require more numerical accuracy than the 15 digits available in double-precision floating variables. The standard approximations for fixation probabilities and sojourn times of mutations, for the respective cases of either neutrality or strong selection, were then used (HALDANE 1924, 1927; KIMURA 1962; KIMURA and OHTA 1969a,b).

To integrate over the distribution of selection coefficients, we partitioned the range of mutational effects of interest, from effectively neutral ( $s = 10^{-10}$ ) to lethal ( $s \geq 1$ ), into  $I$  groups small enough to assume constant mutational effects within each group. We found that  $\pi_A$  and  $K_A$  values computed from  $I = 30, 100,$  and  $300$  equidistant steps on a log scale were accurate to  $\sim 2, 0.2,$  and  $0.02\%$  relative error, respectively. For each bin, we then independently computed (i) the probability  $P_i$  that a mutation will have an effect that belongs to bin  $i$ , obtained by integrating  $\phi(s)$  from the lower to the upper limit of the bin, and (ii) all interesting quantities of the model, given the average mutational effect characterizing that bin. To get the overall result for a parameter of interest, we summed the corresponding results for the parameter over all bins, weighting the value for each bin by  $P_i$ .

Experimenting with different  $\phi(s)$  functions, two special cases became obvious. Some mutations have effects smaller than the lower integration limit ( $s = 10^{-10}$ ). We added the probability mass of these mutations to the first bin. Since these are effectively neutral mutations, this amounts to full integration of the  $\phi(s)$  down to neutrality. The distribution of  $s$  was truncated at  $s \geq 1$ , keeping a record of the fraction of mutations that fall into this category; this represents dominant lethal mutations caused by amino acid substitutions, which are probably extremely rare.

For each quantity involved in the model, a plot over the whole range of values of  $s$  generated from a given  $\phi(s)$  was produced, and the smoothness of transitions to neutrality and to strong selection equilibrium was used to verify the accuracy of the computations. At statistical equilibrium under drift, mutation, and selection at each site, we expect equal rates of substitution between preferred and unpreferred amino acids at sites with a given  $s$ , which was confirmed in our plots.

The fit of the nucleotide site diversity from the two species to a given set of assumed parameters was assessed, using the numerical criterion

$$d = \log_{10} \left( 10^{-50} + \left| \frac{\hat{\pi}_{S_1}}{\hat{\pi}_{A_1}} - \frac{\bar{\pi}_{S_1}}{\bar{\pi}_{A_1}} \right| + \left| \frac{\hat{\pi}_{S_2}}{\hat{\pi}_{A_2}} - \frac{\bar{\pi}_{S_2}}{\bar{\pi}_{A_2}} \right| \right),$$

where  $\hat{\pi}_{A_i}$  is the predicted nonsynonymous diversity value for species 1,  $\bar{\pi}_{A_i}$  is the corresponding observation, and the other corresponding values are for silent diversity and for species 2, using Equation 5a below.

The  $d$  function was chosen to make small differences look large, a property needed for the simplex optimization routine (Amoeba) that we used (NELDER and MEAD 1965), implemented in R. All ratios were rounded to at least six digits to ensure that bad fits were detectable. An optimization attempt was considered as successful if the data could be predicted with six-digit accuracy, using  $I = 100$  integration steps; this excluded cases where optimization stopped without getting close to the data, wrongly suggesting that the data had been fitted. In all cases, the estimates resulting from the first Amoeba run were used as starting values for a second run, to make sure that the first result was not simply a local optimum. The parameter values that satisfied the optimization criterion after the second run were used to compute the results reported here, with an increased accuracy ( $I = 300$ ).

To simplify the calculations using Equations 5 below, we assumed that the effective population size  $N_e$  is equal to the size of the breeding population,  $N$ . To obtain the relevant numerical values, we used Equation 1a below to estimate  $N_e$  from the observed mean silent diversity, assuming a mutation rate  $u$ . We mostly used  $u = 1.5 \times 10^{-9}$  in the calculations, a value widely used for *Drosophila* (POWELL 1997). Since this is not firmly established, analyses with other mutation rates were also done, to check sensitivity to  $u$ .

Once the parameters determining variability from Equations 5a were estimated, they were used in Equation 5b to predict the rate of amino acid substitutions arising from sites under purifying selection, assuming an arbitrary value for the unknown ancestral  $N_e$ . This in turn may be used to estimate the fraction of selectively advantageous amino acid substitutions, by comparing the prediction from Equation 5b with the observed divergence between species, similarly to Equation 4b.

**Statistical analyses:** Preliminary analyses of the data showed that *D. pseudoobscura* had much greater silent diversity than *D. miranda*. Its genes also consistently show an excess of rare variants over neutral expectation (MACHADO *et al.* 2002; SCHAEFFER 2002), in contrast to what is seen in *D. miranda* (YI *et al.* 2003; BARTOLOMÉ *et al.* 2005). This suggests that *D. pseudoobscura* has undergone a recent period of population expansion and is therefore not likely to have reached its final level of neutral or nearly neutral diversity. The measure of neutral variability provided by Watterson's  $\theta_w$  estimator, based on the number of segregating sites for silent diversities (WATTERSON 1975), is probably closer to the equilibrium value than the pairwise nucleotide site diversity estimator  $\pi$ , since new variants arising after an increase in  $N$  are predominantly rare (TAJIMA 1989). We therefore expect  $\theta_w$  to provide a better estimate of the equilibrium neutral/nearly neutral diversity for *D. pseudoobscura* than  $\pi$  and have accordingly used  $\theta_w$  for silent sites for both species. For sites under selection, there is no cogent reason to use  $\theta_w$ , and so we used  $\pi$  as the diversity measure for nonsynonymous variants. We obtained very similar results if  $\theta_w$  is used for both types of sites. For simplicity, we use the symbol  $\pi$  to denote diversity estimates for both cases in what follows.

Mean values across genes of the diversities and divergence statistics were used to estimate the parameters of the models. Weighted mean diversity estimates were obtained using the inverse of the sum of the estimated sampling and stochastic

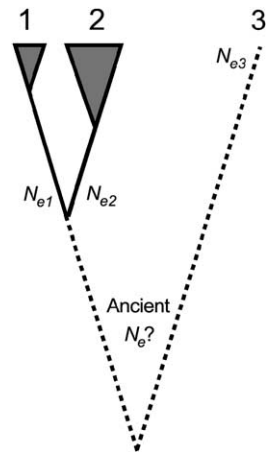


FIGURE 1.—Basic three-species setting. We assume fixed values for all three effective population sizes  $N_e$ , despite uncertainty concerning the sizes of ancestral populations. The shaded areas indicate the availability of neutral and selected polymorphism data from the two more closely related species with different  $N_e$ . The dotted line to the third species indicates that our method for inferring the strength of purifying selection also applies to cases where only polymorphism data for two species exist.

evolutionary variances of diversity as the weight for a given gene, obtained from the standard formulas under neutrality and free recombination (different formulas apply to the Watterson and pairwise diversity estimators as described in Chapter 10 of NEI 1987). This procedure is heuristic, given that there is some linkage disequilibrium among sites within genes and that nonsynonymous mutations are known to be subject to selection, but it provides a simple approximate way of accounting for different levels of noise across genes. Divergence values,  $K$ , were weighted by the number of sites involved. Their means across genes were then used to estimate  $K_A/K_S$ . For comparison, unweighted means of diversities and divergence were also computed and used for parameter estimation.

To assess the variability of our estimates for both methods of weighting, we generated 1000 “observations” by bootstrapping the diversity and divergence data across loci, as described by BARTOLOMÉ *et al.* (2005). The upper and lower fifth percentiles of the distribution of each parameter were used as approximate 90% confidence intervals for the parameters in question; this provides a convenient basis for assessing 5%  $P$ -values for the null hypothesis that this parameter has a value of zero, in a one-tailed test.

## THEORETICAL FRAMEWORK

**General framework:** We assume that the sequences of population samples of many different genes are known, giving reliable estimates of diversities for nonsynonymous and silent or synonymous sites in each of two species and corresponding divergences from a third (Figure 1). Differences in  $N_e$  exist between the two species for which diversity data are available; the effective population size of species  $i$  ( $i = 1$  or  $2$ ) is denoted by  $N_{e,i}$ , where the smaller  $i$  denotes the species with the smaller  $N_e$ . We assume an infinite-sites model with autosomal

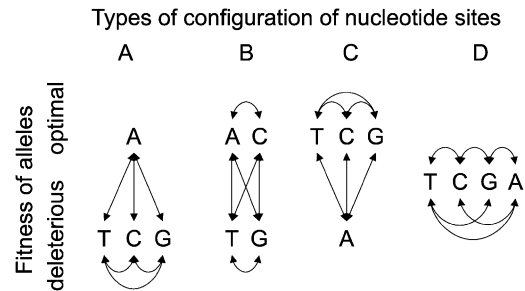


FIGURE 2.—All possible fitness-relevant configurations for nondegenerate nucleotide sites. The four alternatives A, T, C, and G stand for any of the four possible nucleotides at a site. In configuration A, only one allele is optimal, whereas in D all alternatives are equally fit (*i.e.*, they are mutually neutral). B and C allow for neutral mutations between equally optimal alternatives at sites that are capable of producing deleterious mutations. This study assumes that configurations B and C do not exist and estimates the frequency of configuration D from the data ( $c_n$ ). This will lead to upper limits on estimates of the frequencies of configurations A and D. As long as no further mutations occur at segregating sites (the infinite-sites assumption), all four possible alternatives can be collapsed to the specific case of two alleles described in the text.

inheritance (KIMURA 1971). Silent mutations are assumed to be selectively neutral or sufficiently close to neutrality that their evolutionary fates are well described by the neutral model. Each site is assumed to evolve independently; *i.e.*, recombination is sufficiently frequent that Hill-Robertson interference effects can be neglected. Mutation rates and selection coefficients at each site are assumed to be independent, and random mating is assumed for both species.

To keep our model simple, we assume that there are at most two types of amino acid at a given nucleotide site: “preferred” and “unpreferred,” with at most one of the four possible nucleotides corresponding to the preferred state (our methods do not require us to identify preferred and unpreferred states from the data). At some sites, all possible variants may be effectively neutral, meaning in practice that variants are subject to a similar level of selection to silent mutations (Figure 2D). The fraction of such “neutral” nonsynonymous mutations is denoted by  $c_n$ . We treat other sites as in Figure 2A, where nonsynonymous nucleotide site mutations can result in a change to a selectively deleterious amino acid (if the nucleotide in question codes for the preferred state), to a favorable amino acid (if the site is fixed for an unpreferred amino acid, and the mutation leads back to the preferred state), or else to a selectively neutral variant (if the site is fixed for an unpreferred amino acid, and the mutation causes a change to a different but selectively equivalent deleterious variant). Because some mutations at such sites are deleterious, we refer to these sites as undergoing purifying selection.

The selection coefficient  $s$  for selectively deleterious variants refers to the reduction in fitness of heterozygotes; this may vary according to the site and amino

acid in question. For convenience of calculation, dominance is assumed to be intermediate. As in standard models of codon usage bias (LI 1987; BULMER 1991; McVEAN and CHARLESWORTH 1999), if selection at these sites is sufficiently weak, unpreferred mutations can become fixed and then mutate back to the preferred state, contributing to a flux of amino acid substitutions as well as to polymorphism. This model is not completely general (see Figure 2, B and C), but should suffice as a guide to the basic processes involved, especially when selection is fairly strong in relation to drift.

Occasionally, a new adaptive mutation (distinct from a back mutation at a site fixed for a deleterious mutation) may arise at a nonsynonymous site, perhaps in response to a change in the environment. This is assumed to spread rapidly to fixation if it survives initial stochastic loss. The substitution rate per generation per site for mutations that fall into this category is denoted by  $c_a u$ , where  $u$  is the expected mutation rate per nucleotide site, and  $c_a$  measures the substitution rate as a fraction of all mutations.  $c_a$  is the product of the frequency of adaptive mutations and their fixation probability, integrated over all advantageous effects. We assume that  $c_a \ll 1$ , since favorable mutations are likely to be rare.

These assumptions lead to the following general equations for the expectations of the silent and nonsynonymous nucleotide site diversities in species  $i$ ,  $\pi_{S_i}$  and  $\pi_{A_i}$ , and the rates of substitution per site per generation for silent and nonsynonymous mutations,  $K_{S_i}$  and  $K_{A_i}$  (second-order terms in small quantities are ignored):

$$\pi_{S_i} = 4N_{e_i} u \tag{1a}$$

$$\pi_{A_i} = 4c_n N_{e_i} u + (1 - c_n) H_{P_i} \tag{1b}$$

$$K_{S_i} = u \tag{1c}$$

$$K_{A_i} = c_n u + (1 - c_n) K_{P_i} + c_a u, \tag{1d}$$

where  $H_{P_i}$  is the mean equilibrium diversity at sites subject to purifying selection, and  $K_{P_i}$  is the mean substitution rate at such sites.

**Strong purifying selection:** These equations become greatly simplified if  $N_e s > 1$  for all nonneutral nonsynonymous mutations in both species. The equilibrium diversity contributed by sites subject to purifying selection with selection coefficient  $s$  is then well approximated by the deterministic expression,  $2u/s$ , as can be shown by numerical solutions of the general equations (McVEAN and CHARLESWORTH 1999). We then have

$$\pi_{A_i} = c_n \theta_i + 2(1 - c_n) \frac{u}{s_h}, \tag{2a}$$

where  $\theta_i = 4N_{e_i} u$ , and  $s_h$  is the harmonic mean of the selection coefficients for all mutations that are not effectively neutral (this is the same for both species, since we assume that nearly neutral mutations are absent).

For  $N_e s > 2$ ,  $K_{P_i}$  is negligibly small, so that we can replace Equation 1d by

$$K_{A_i} = c_n u + c_a u. \tag{2b}$$

With just two species for which diversity data are available, Equations 1a and 2a lead to the following formula for  $c_n$ :

$$c_n = \frac{(\pi_{A_2} - \pi_{A_1})}{(\pi_{S_2} - \pi_{S_1})}. \tag{3a}$$

Substituting this into Equation 2a and using Equation 1a, we obtain

$$2N_{e_1} s_h = \frac{\pi_{S_1} (\pi_{A_1} + \pi_{S_2} - \pi_{A_2} - \pi_{S_1})}{\{\pi_{A_1} (\pi_{S_2} - \pi_{S_1}) - \pi_{S_1} (\pi_{A_2} - \pi_{A_1})\}}. \tag{3b}$$

From Equations 1c and 2b,  $c_a$  is given by

$$c_a = \frac{K_A}{K_S} - c_n, \tag{4a}$$

assuming that species 1 and 2 have the same mean divergences for silent and nonsynonymous sites from the third species, so that subscripts can be dropped. This is necessarily the case with strong selection, since species 1 and 2 are equally close to species 3, and only neutral nonsynonymous mutations can become fixed (there is no reason in principle why divergence between species 1 and 2 could not be used in the absence of data on a third species, but in the present case the level of divergence between *D. miranda* and *D. pseudoboscuro* is so low that estimates based on this would be very unreliable).

The proportion of nonsynonymous substitutions that are caused by the fixation of advantageous mutations is

$$P_a = \frac{K_S c_a}{K_A}. \tag{4b}$$

Given the above assumptions, all the parameters of the model can be estimated by equating expectations to observed values.

**Arbitrary purifying selection:** The validity of the assumption that  $N_e s > 1$  for all nonneutral nonsynonymous mutations is, however, questionable. If this assumption is relaxed, the formulas for the equilibrium diversity at nonsynonymous sites become more complex, and the possibility of a contribution to  $K_A$  from sites subject to purifying selection must also be considered, using a probability distribution  $\phi(s)$  of selection coefficients for mutations subject to selection. We now consider this problem in detail.

As far as diversity is concerned, we note first that a nonsynonymous site that has become fixed for a mutant nucleotide coding for a deleterious amino acid can either mutate back to the original nucleotide or mutate to another nucleotide coding for a deleterious amino acid. If mutation rates among all four nucleotides were

equal, the probability of back mutation to the original state would be  $\frac{1}{3}$ ; in general, however, inequalities in mutation rates are likely to make this probability different from  $\frac{1}{3}$ , and so we represent it as  $1/\kappa$ .

$\kappa$  is the ratio of the forward and backward mutation rates for mutations creating deleterious amino acids, a measure similar to the mutational bias parameter used in models of codon usage bias (LI 1987; BULMER 1991; McVEAN and CHARLESWORTH 1999). At a site fixed for a deleterious amino acid, there will thus be a mutation rate  $u/\kappa$  back to the preferred amino acid; there will also be a neutral mutation rate  $u(\kappa - 1)/\kappa$ , if all deleterious amino acids at this site are selectively equivalent. These should be taken into account in the contribution to net diversity. Using the argument that led to Equations 6 and 7 of McVEAN and CHARLESWORTH (1999), we then find that  $H_{P_i}$  in Equation 1b is given by

$$H_{P_i} = 2N_i u \int \left\{ \left( \frac{2(N_{e_i}/N_i)(\kappa - 1) + H_0(s)}{\kappa} \right) \times m_0(s) + H_1(s) m_1(s) \right\} \phi(s) ds, \quad (5a)$$

where  $N_i$  is the total number of breeding individuals in species  $i$ ;  $H_0(s)$  and  $H_1(s)$  are the expected total heterozygosities that are contributed during their sojourn in the population by new mutations to preferred and unpreferred amino acids, respectively, for a selection coefficient  $s$ ;  $m_0(s)$  and  $m_1(s)$  are the fractions of sites fixed for unpreferred and preferred amino acids, respectively, among sites with selection coefficient  $s$ ; and  $\phi(s)$  is the probability density function for the distribution of selection coefficients. Formulas for the  $H$  and  $m$  functions are given by McVEAN and CHARLESWORTH (1999), Equations 5 and 10.

Similarly,  $K_{P_i}$  in Equation 1d is given by

$$K_{P_i} = 2N_i u \int \left\{ \left( \frac{(\kappa - 1)/(2N_i) + U_0(s)}{\kappa} \right) \times m_0(s) + U_1(s) m_1(s) \right\} \phi(s) ds, \quad (5b)$$

where  $U_0(s)$  and  $U_1(s)$  are the fixation probabilities for new mutations to preferred and unpreferred amino acids, respectively, given a selection coefficient  $s$ , using the standard diffusion equation formula (KIMURA 1962).

Even if the distribution  $\phi(s)$  is described by only two parameters, such as the arithmetic mean and standard deviation, there are too few degrees of freedom in the data to estimate all the parameters of interest by equating observed and expected values of diversities and divergence, unless we are prepared to assume that there are no nonsynonymous sites with neutral effects ( $c_n = 0$ ). These equations do, however, provide a means of evaluating the sensitivity of the results to our assumptions about  $c_n$  or the properties of the distribution, as is described below. Following convention (PIGANEAU and

EYRE-WALKER 2003), we use the gamma distribution for  $\phi(s)$ ,

$$\phi(s) = \frac{s^{\alpha-1} \exp(-s/\beta)}{\beta^\alpha \Gamma(\alpha)}, \quad (6)$$

where  $\alpha$  is the shape parameter,  $\beta$  is the scale parameter,  $\Gamma$  is the gamma function,  $\alpha\beta$  is the arithmetic mean, and  $\alpha\beta^2$  is the variance of  $s$  (R-PROJECT 2005).

## RESULTS

We now present the results of applying these methods to the diversity data on *D. miranda* and *D. pseudoobscura* described in MATERIALS AND METHODS (*D. miranda* is designated as species 1 and *D. pseudoobscura* as species 2 in what follows). In view of such problems as the lack of overlap between the genes used in the two species, and the disparity in sample sizes between studies, the results based on these data should be regarded as merely provisional and illustrative of the methods.

**Strong purifying selection:** We first present the results of applying the expectations for the case of strong purifying selection. Divergence values from *D. affinis* (species 3) were estimated for the 17 genes surveyed in *D. miranda*; the divergence values for these genes between *D. affinis* and *D. pseudoobscura* were almost identical (BARTOLOMÉ *et al.* 2005), so that only the *D. miranda* results were used here. Both weighted and unweighted estimates of mean diversities were obtained, as described in MATERIALS AND METHODS; statistical uncertainty was assessed by bootstrapping across genes.

The results are displayed in Table 1. The main conclusion is that an estimate of  $N_{e_i,sh}$  substantially greater than one is supported by the bootstrapping results, even using the unweighted estimators, which yield lower values than the weighted estimators. The distribution of  $N_{e_i,sh}$  is, however, very wide, and infinite values were sometimes generated for the weighted data. The estimate of  $N_{e_i,sh}$  for *D. pseudoobscura* was, of course, proportionately larger, corresponding to the large  $N_e$ -value estimated from silent site diversities (these suggested a 5.8-fold higher  $N_e$  for *D. pseudoobscura*).

The estimated proportion of neutral sites was smaller with the unweighted estimates than with the weighted ones, but both suggested a value of a few percent. There was a very wide distribution of values of the proportion of fixations due to adaptive mutations in both cases, and a zero value could not be ruled out by the data, although the value estimated from the data exceeded 50% for the unweighted estimate.

**Arbitrary purifying selection:** We now relax the assumption that  $\pi_A$  is close to the deterministic prediction by applying the model behind Equations 5 to the same data. The assumptions of a gamma distribution of mutational effects, and that 0, 2.5, or 5% of all nonsynonymous mutations are neutral, yielded the parameters reported in Tables 2 and 3 for the weighted and

**TABLE 1**  
**Summary of the data and estimates from the strong purifying selection model**

Estimate	$\pi_{A_1}$	$\theta_{S_1}$	$\pi_{A_2}$	$\theta_{S_2}$	$K_A$	$K_S$	$N_{e_1,sh}$	$c_n$	$P_a$
Weighted	0.086 (0.041/0.136)	0.502 (0.340/0.710)	0.294 (0.196/0.400)	2.86 (2.02/3.43)	2.11 (1.28/2.80)	23.0 (20.4/24.9)	5.41 (2.84/∞)	8.79 (3.91/16.5)	3.99 (-123/57.5)
Unweighted	0.088 (0.044/0.141)	0.478 (0.342/0.626)	0.206 (0.124/0.300)	2.73 (2.31/3.14)	2.48 (1.30/3.76)	22.2 (19.9/24.8)	3.58 (1.80/29.8)	5.24 (0.923/10.3)	52.9 (-28.9/93.3)

All values except for  $N_{e_1,sh}$  are expressed as percentages. An “infinite” value of  $N_{e_1,sh}$  corresponds to a zero or negative denominator of Equation 3b. Values in parentheses give the approximate lower and upper fifth percentiles from 1000 bootstrap replicates (see text for details).  $\pi_{A_i}$  is the nonsynonymous diversity for species  $i$  ( $i = 1$  for *D. miranda* and 2 for *D. pseudoobscura*),  $\theta_{S_i}$  is the silent diversity for species  $i$ ,  $K_A$  and  $K_S$  are the nonsynonymous and silent divergences between *D. miranda* and *D. affinis*,  $c_n$  is the fraction of completely neutral mutations, and  $P_a$  is the proportion of adaptive nonsynonymous mutations.

unweighted data, respectively. With values of  $c_n \geq 7.5\%$  we rarely, if ever, found parameters for a gamma distribution that fitted the data, and the fraction of bad fits for  $c_n = 5\%$  was 48.5% for 1000 bootstraps from the weighted data, whereas the mean of the unweighted data could not be fitted assuming  $c_n \geq 5\%$ . With the gamma distribution, a significant fraction of mutations had such small selection coefficients that  $N_e s < 1$  or even 0.5. The results of McVEAN and CHARLESWORTH (1999) suggest that, with  $N_e s < 0.5$ , both diversities and substitution rates are nearly equivalent to those for neutral sites; in addition, the intensity of selection on synonymous mutations that change codon usage from preferred to unpreferred codons in *D. miranda* seems to be close to  $N_e s = 0.5$  (BARTOLOMÉ *et al.* 2005). It thus seems reasonable to use this value as the boundary for designating mutations as effectively neutral; the sum of  $c_n$  and the fraction of effectively neutral mutations generated by the fitted gamma distribution is denoted by  $c_{ne}$  in Tables 2 and 3. For completeness, we also show results using  $N_e s = 1$  as the boundary.

Despite the time-consuming nature of the multiple integrations involved in implementing this model, we attempted 1000 bootstrap replications to assess the reliability of our parameter estimates. Out of these bootstrap computations for the weighted or unweighted data in Table 2 or 3, only 80.3 and 90.1%, respectively, could be fitted assuming  $c_n = 0\%$ ; 75.6 and 65.6% could be fitted assuming  $c_n = 2.5\%$ . Technically, our estimates of the distributions of mutational effects involve only the shape and the location parameter for the gamma distribution of  $s$ , together with values for  $N_e$ . However, in Tables 2 and 3 we also report more intuitively meaningful measures of the underlying distribution of all selection coefficients at sites capable of mutating to nonlethal mutations that fall above the threshold of effective neutrality, scaled by  $N_e$ . These include the arithmetic and harmonic means, the coefficient of variation, and the lower and upper fifth percentiles of the distribution of  $s$ -values for these mutations.

As reviewed in more detail in the DISCUSSION, the bulk of the nonneutral mutations segregating in the

population come from the more weakly selected tail of the distribution (SUNYAEV *et al.* 2001). The arithmetic mean of the distribution of selection coefficients among these polymorphic variants is much closer to the harmonic than to the arithmetic mean of the distribution for new mutations, and so the harmonic means in Tables 2 and 3 are more relevant than the arithmetic means to the properties of mutations found in populations. It is notable from Tables 2 and 3 that the means for *D. pseudoobscura* were smaller than might be expected from the ratio of effective population sizes and the corresponding means for *D. miranda*; this reflects the lower fraction of mutations that fall into the effectively neutral class when the effective population size is larger, reducing the average selection coefficients for mutations in the other class.

Given an ancestral  $N_e$ -value, we can also predict the substitution rate for mutations under purifying selection. This can be compared with the observed rate to estimate the proportion of adaptive substitutions, using Equation 4. We report three alternative values:  $P_{a_1}$ ,  $P_{a_2}$ , and  $P_{a_3}$ , which assume ancestral  $N_e$ -values equal to the estimated current  $N_e$  for *D. miranda*, the current  $N_e$  of *D. pseudoobscura*, and the mean of these, respectively. For each value, we give the approximate lower and upper fifth percentiles obtained by bootstrapping.

The main result of Tables 2 and 3 is solid support for the conclusion that ~90% or more of all amino acid mutations have significantly deleterious effects. It is also remarkable that the estimates of the biologically important harmonic mean selection coefficient are close to those using the strong selection assumption, taking into account the statistical noise and the uncertainty regarding  $c_n$  for the arbitrary purifying selection model. Similar conclusions can be drawn for comparisons between  $c_{ne}$  in the arbitrary purifying selection model and  $c_n$  in the strong selection model. Comparing the weighted and unweighted estimates shows that the weighting procedure influences the estimates, but the noise in the data is larger than the noise from uncertainty about the best weighting procedure. While our definitions of effective neutrality ( $N_e s < 0.5$  or  $< 1$ ) have

**TABLE 2**  
**Estimates of the distribution of mutational effects for variance-weighted data from *D. miranda* and *D. pseudoobscura***

$c_n$ (%)	$\alpha$	loc	Species	$N_{e,s}$ (am)	$N_{e,s}$ (hm)	$N_{e,s}$ (5%)	$N_{e,s}$ (95%)	CV	$c_{ne}$ (%)	$P_{a_1}; P_{a_2}$ (%)	$P_{a_3}$ (%)	
0	0.299 (0.0782/0.741)	0.000266 ( $1.14 \times 10^{-5}$ /17,000)	<i>mir</i>	255	8.35	1.31	1,100	1.67	12.7	-42.8	0.902	
				(11.6/95,700)	(3.49/17.9)	(0.843/3.01)	(36.9/558,000)	(1.08/2.15)	(7.91/17.0)	(-125/14.7)	(-83.5/60.3)	
			<i>pso</i>	263	13.3	2.22	1,100	1.64	15.5			
				(12.4/97,500)	(4.82/31.7)	(1.40/5.06)	(37.8/560,000)	(1.02/2.11)	(11.0/20.2)			
				1,370	14.4	2.50	6,220	1.74	7.57	15.7		
				(55.3/521,000)	(8.69/25.0)	(1.55/5.47)	(188/3,110,000)	(1.14/2.22)	(2.35/13.4)	(-72.1/73.7)		
1,400	23.0	3.72	6,240	1.72	9.19							
(56.2/534,000)	(11.8/43.8)	(2.39/8.47)	(188/3,200,000)	(1.12/2.20)	(3.65/14.7)							
2.5	0.448 (0.0996/1.32)	$7.67 \times 10^{-5}$ ( $7.12 \times 10^{-6}$ /89.5)	<i>mir</i>	70.6	6.86	1.21	274	1.39	11.4	-31.1	7.93	
				(6.96/76,500)	(3.19/17.2)	(0.877/2.96)	(18.4/463,000)	(0.832/2.08)	(5.73/16.6)	(-117/27.9)	(-78.7/57.7)	
			<i>pso</i>	73.1	10.1	1.93	275	1.36	14.4			
				(7.32/78,800)	(3.99/29.9)	(1.37/4.80)	(18.8/476,000)	(0.789/2.05)	(9.04/19.3)			
				382	14.7	2.73	1,550	1.45	6.58	16.4		
				(35.4/392,000)	(9.48/28.4)	(1.77/7.31)	(97.2/2,450,000)	(0.865/2.13)	(2.87/13.4)	(-70.0/59.1)		
388	21.5	3.96	1,550	1.43	7.95							
(35.4/401,000)	(12.2/45.5)	(2.59/9.77)	(97.2/2,490,000)	(0.857/2.12)	(3.30/14.6)							
5	0.831 (0.228/1.75)	$2.33 \times 10^{-5}$ ( $6.84 \times 10^{-6}$ /0.0179)	<i>mir</i>	20.4	5.42	1.21	63.6	1.05	9.27	-11.2	15.2	
				(6.53/11,400)	(3.39/18.9)	(0.931/3.53)	(15.8/55,700)	(0.737/1.95)	(6.40/16.1)	(-113/22.6)	(-78.3/38.3)	
			<i>pso</i>	21.1	7.07	1.76	63.7	1.02	12.2			
				(6.72/11,600)	(4.08/31.5)	(1.42/5.40)	(16.2/55,800)	(0.702/1.93)	(8.41/19.0)			
				112	17.4	4.28	362	1.09	6.01	14.4		
				(31.7/80,400)	(10.5/35.2)	(2.13/10.9)	(80.1/413,000)	(0.753/1.98)	(5.08/12.1)	(-73.1/35.6)		
113	21.5	4.73	362	1.08	6.72							
(32.0/81,200)	(12.5/53.0)	(2.93/13.1)	(80.2/415,000)	(0.746/1.97)	(5.22/13.7)							

The  $N_{e,s}$  values reported here describe the distribution of deleterious mutational effects for all effectively nonneutral and nonlethal sites, where (am) denotes the arithmetic mean and (hm) the harmonic mean; (5%) and (95%) represent the lower and upper fifth percentiles of this truncated distribution; *mir* and *pso* refer to *D. miranda* and *D. pseudoobscura*, respectively;  $c_n$  is the fraction of completely neutral mutations assumed in the calculations;  $c_{ne}$  is the fraction of effectively neutral mutations; CV denotes the coefficient of variation of  $N_{e,s}$  values of the truncated distribution;  $P_{a_1}$ ,  $P_{a_2}$ , and  $P_{a_3}$  are the proportions of adaptive amino acid substitutions, assuming either the current population size of *D. miranda* or *D. pseudoobscura* or their mean as the ancestral  $N_e$ , respectively. For a given species and value of  $c_n$ , the upper and lower values for all estimates of  $N_{e,s}$ ,  $c_{ne}$ , and CV correspond to assumed borders to neutrality of  $N_{e,s} = 0.5$  and 1, respectively. Values in brackets indicate the lower and upper approximate fifth percentiles from bootstrapping across genes, after omitting all values that did not fit the data (see text). All estimates assume a mutational bias of  $\kappa = 2$ , a mutation rate of  $\mu = 1.5 \times 10^{-9}$ , and a gamma distribution of mutational effects with shape parameter  $\alpha$  and location parameter  $\beta$ , where  $\beta$  is computed from loc, the arithmetic mean of the untruncated distribution of selection coefficients given in the table (see text).

TABLE 3  
Estimates of the distribution of mutational effects for unweighted data from *D. miranda* and *D. pseudoobscura*

$c_n$ (%)	$\alpha$	loc	Species	$N_{e,s}$ (am)	$N_{e,s}$ (hm)	$N_{e,s}$ (5%)	$N_{e,s}$ (95%)	CV	$c_{ne}$ (%)	$P_{a1}; P_{a2}$ (%)	$P_{a3}$ (%)
0	0.562 (0.172/1.62)	$2.75 \times 10^{-5}$ $(4.25 \times 10^{-6}/0.0732)$	<i>mir</i>	24.3	4.76	0.978	87.6	1.23	9.93	4.03	51.5
				(3.89/62,700)	(2.20/16.4)	(0.730/2.67)	(9.76/347,000)	(0.735/1.96)	(4.43/14.4)	(-61.3/49.9)	(-12.7/91.1)
				25.6	6.77	1.56	88.5	1.18	14.2		
			<i>pso</i>	(4.21/63,700)	(2.83/29.1)	(1.19/4.52)	(9.84/348,000)	(0.666/1.93)	(8.73/20.9)	64.2	
				(19.7/320,000)	(7.72/24.3)	(1.65/5.49)	(51.4/1,760,000)	(0.784/2.02)	(0.332/8.86)	(-2.18/96.5)	
				133	16.1	3.02	472	1.28	5.33		
2.5	1.05 (0.267/2.48)	$1.20 \times 10^{-5}$ $(4.92 \times 10^{-6}/0.00238)$	<i>mir</i>	(19.8/324,000)	(9.39/41.3)	(2.28/7.80)	(51.5/1,770,000)	(0.780/2.00)	(0.965/9.99)	25.4	58.4
				(3.91/2,050)	(2.41/15.6)	(0.786/2.63)	(8.97/9,060)	(0.624/1.80)	(3.53/13.4)	(-48.1/57.8)	(-11.7/73.6)
				10.0	3.87	0.987	29.8	0.930	7.13		
			<i>pso</i>	10.5	4.95	1.46	30.0	0.887	11.3	61.2	
				(4.11/2,070)	(2.94/24.8)	(1.23/4.07)	(9.06/9,110)	(0.591/1.78)	(5.76/18.3)		
				55.2	14.0	3.53	161	0.968	3.25		
				(21.9/11,300)	(8.81/29.2)	(2.07/8.57)	(50.1/50,700)	(0.634/1.84)	(2.52/8.07)	(-5.50/73.7)	
				55.6	16.3	4.09	161	0.960	3.95		
				(22.2/11,300)	(10.7/43.9)	(2.71/9.88)	(50.8/50,700)	(0.633/1.83)	(2.58/9.58)		

See Table 2 for explanation.

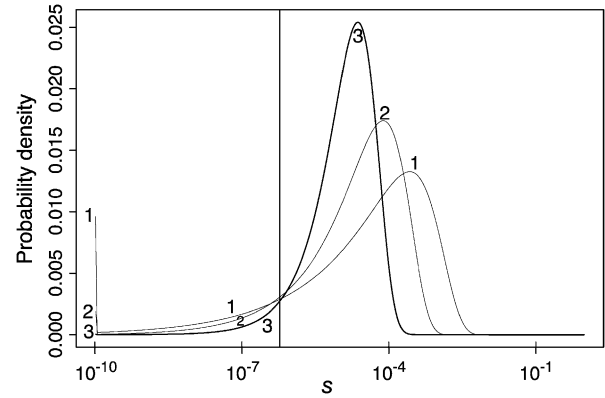


FIGURE 3.—Gamma distributions of mutational effects estimated from the variance-weighted data, assuming  $c_n = 0, 2.5,$  and  $5\%$  (curves 1, 2, and 3, respectively). The spike below  $s = 10^{-10}$  corresponds to the integral of the distribution from 0 up to this value. The vertical line corresponds to  $N_e s = 0.5$  for *D. miranda*. The computations assume  $u = 1.5 \times 10^{-9}$  and  $\kappa = 2$ . Note the use of a log scale for the selection coefficients.

little influence on the arithmetic mean and the upper fifth percentile of the distribution of mutational effects, their influence on the harmonic mean and the lower fifth percentile is greater. In no case, however, did the  $\sim 1\text{--}3\%$  of mutations that fall in the range  $0.5 < N_e s < 1$  change our conclusions regarding the prevalence of deleterious mutations.

To visualize the gamma distributions of mutational effects estimated from the weighted data, we plotted our best-fitting estimates with assumed values of  $c_n = 0, 2.5,$  and  $5\%$  (Figure 3). This involved estimating effective population size from Equation 1a, using the “standard” mutation rate of  $1.5 \times 10^{-9}$ . The distributions for the two smaller values of  $c_n$  peaked at selection coefficients  $\sim 10^{-4}$ , fairly close to the arithmetic means obtained from Table 2 for mutations that are not effectively neutral, whereas the harmonic mean was about one-tenth of this.

## DISCUSSION

### Robustness of the strong purifying selection model:

We found quite good agreement between the results of the strong purifying selection model and the model with a gamma distribution of selection coefficients for the estimates of the harmonic mean of the selection coefficient. This suggests that estimates of the magnitude of this important parameter are robust to the assumptions used, providing that a sizeable fraction of nonsynonymous polymorphisms is nonneutral. Some approximations that relax the assumptions of the strong purifying selection model, but that do not depend on the details of the distribution, are explored in the APPENDIX. These provide very general methods for estimating  $N_e s_{hm}$  and again suggest that the magnitudes of the estimates of the main parameters of interest are fairly robust to details of the assumptions.

To simplify even further, one could assume that all polymorphic nonsynonymous sites are effectively under purifying selection; *i.e.*,  $c_n = 0$ . The mean frequency of such deleterious mutations is given by  $q = u/s_h$  (see Equation 2a), where  $u$  is the mutation rate per site per generation and  $s_h$  is the harmonic mean of the heterozygous selection coefficients. Since  $q$  is small, we have

$$s_h \approx 2u/\pi_A. \quad (8a)$$

Using  $\pi_S = 4N_e u$ , we obtain

$$N_e s_h \approx \pi_S/(2\pi_A). \quad (8b)$$

This result is remarkably robust, since we do not need to know  $N_e$ ,  $u$ ,  $s_h$ , or dominance coefficients. However, caution is necessary if Equation 8b gives values near 1, since this suggests that drift is probably too strong to be neglected. In this case, the true strength of selection for the deleterious mutations will be larger than predicted by this approach. The values estimated from this method are  $N_e s_h \approx 2.9$  and  $4.9$  for *D. miranda* and *D. pseudoobscura*, as estimated for the respective weighted data. These values are little more than a third of the respective values estimated from the most precise methods and lie outside the confidence intervals for the latter. This very simple formula is therefore too crude for precise estimates, but seems to work reasonably well as a rough first estimate for the lower bound of  $N_e s_h$ . It can be applied only when there is evidence that a substantial proportion of segregating nonsynonymous variants experience purifying selection, as in the present case (BARTOLOMÉ *et al.* 2005).

One caveat should be noted concerning the estimates of  $N_e$  for *D. pseudoobscura* that we have been using. This assumes that silent sites are effectively neutral, which we have taken to mean that  $N_e s$  is of the order of  $\leq 0.5$  McVEAN and CHARLESWORTH (1999). While *D. miranda* synonymous sites seem to satisfy this condition (BARTOLOMÉ *et al.* 2005), this condition clearly cannot apply to *D. pseudoobscura*, given that mean silent diversity is four to five times higher than that in *D. miranda*, unless selection on synonymous sites is much weaker in *D. pseudoobscura*. AKASHI and SCHAEFFER (1997) estimated  $N_e s$  against unpreferred codons to be 4.6 (95% confidence interval 2.4–12.1) for *Adh* plus *Adhr* in *D. pseudoobscura*, (although selection for preferred codons was negligible); this result may be in part confounded by the effects of population expansion. If  $N_e s$  for synonymous sites in *D. pseudoobscura* is indeed higher than that in *D. miranda*, then mean silent site diversity (which includes a large contribution from synonymous sites) will yield an underestimate of  $4N_e u$  for this species. Accordingly,  $N_e s$  for nonsynonymous sites will be underestimated and the proportion of effectively neutral nonsynonymous mutations overestimated, since the ratio of nonsynonymous diversity relative to effectively neutral diversity will be overestimated for this species.

TABLE 4

The influence of mutation rate  $u$  and mutational bias  $\kappa$  on estimates of  $c_{ne}$ ,  $N_e s_h$  and  $\alpha$

	$u = 0.5 \times 10^{-9}$	$u = 2 \times 10^{-9}$	$u = 8 \times 10^{-9}$
$\kappa = 1$	10.8%	10.9%	11.0%
	6.51	6.59	6.67
	0.486	0.477	0.469
$\kappa = 2$	11.3%	11.4%	11.5%
	6.88	6.95	7.03
	0.453	0.447	0.441
$\kappa = 3$	11.4%	11.5%	11.6%
	6.95	7.02	7.10
	0.452	0.446	0.440

The top value in each row gives  $c_{ne}$  (in percent). The middle value gives  $N_e s_h$  [comparable to  $N_e s$  (hm) in Table 2]. The bottom value gives  $\alpha$ , the estimated shape of the gamma distribution. All values are for variance-weighted data from *D. miranda*, assuming  $c_n = 2.5\%$  and  $N_e s = 0.5$  as the border of neutrality.

**Sensitivity to mutation rates, mutational bias, and recombination rates:** Estimates derived from the arbitrary purifying selection model are surprisingly insensitive to plausible changes in mutational bias and mutation rate. DRAKE *et al.* (1998, p. 1673) estimated from laboratory experiments that  $8.5 \times 10^{-9}$  mutations per base per generation happen in *D. melanogaster*. POWELL (1997, p.369–371) reported rates between  $0.67 \times 10^{-9}$  and  $3.3 \times 10^{-9}$ , estimated from divergence between the *D. melanogaster* and *D. obscura* groups, assuming that divergence happened 30 MYR ago and that each year represents  $\sim 10$  generations. McVEAN and VIEIRA (2001) estimated a rate of  $1.5 \times 10^{-9}$  (95% C.I. =  $1.0 \times 10^{-9}$ – $2.5 \times 10^{-9}$ ), assuming 2–4 MYR divergence between *D. melanogaster* and *D. simulans*. Others (ANDOLFATTO and PRZEWORSKI 2000; PRZEWORSKI *et al.* 2001) found rates of between  $0.6 \times 10^{-9}$  and  $4.75 \times 10^{-9}$ . Thus  $0.5 \times 10^{-9}$  and  $8 \times 10^{-9}$  appear to be reasonable choices for the most extreme lower and upper credible limits for mutation rates in *Drosophila*. As can be seen in Table 4, when  $c_n = 2.5\%$ , a mutational bias of  $\kappa = 1$  leads to the largest differences from our other estimates, while large changes in mutation rate seem to have only minor effects. Other assumed values of  $c_n$  yield the same conclusion, although the resulting estimates differ because of the strong influence of  $c_n$  (see Tables 2 and 3).

Three genes in our *D. pseudoobscura* data set (*Amy1*, *eve*, and *exu1*) are located on Muller's C, a genomic region that is segregating for paracentric inversions (DOBZHANSKY and POWELL 1975) and therefore has a highly reduced recombination rate. These three genes violate the assumption of independence of sites much more than the other genes. We ran 500 bootstraps for a variance weighted data set without these genes under the assumption of  $c_n = 0\%$ . Results indicate that the inclusion of these genes does not strongly affect our

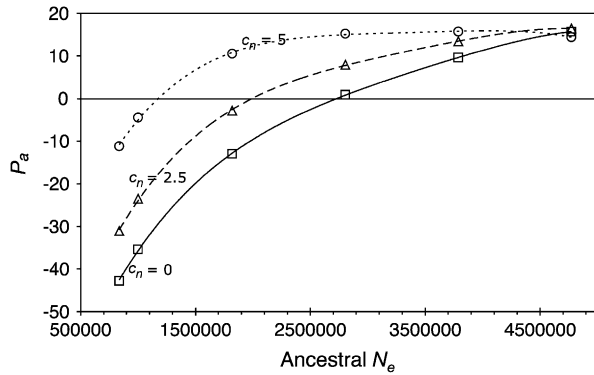


FIGURE 4.—Dependence of estimates of the proportion of adaptive mutations ( $P_a$ ) on ancestral  $N_e$ . All values were computed from the variance-weighted data and assume  $u = 1.5 \times 10^{-9}$ ;  $\kappa = 2$ ; and  $c_n = 0, 2.5$ , and  $5\%$  for the solid, dashed, and dotted curves, respectively. The smallest and largest values of  $N_e$  correspond to the estimates of current  $N_e$  for *D. miranda* and *D. pseudoobscura*, respectively.

parameter estimates, as the confidence intervals of our estimates for the reduced data set mostly overlap our estimates for the full data set (data not shown).

**The reliability of estimates of the proportion of adaptive mutations:** As can be seen from Figure 4, there is a large influence of the value of the (unknown) ancestral  $N_e$  on the estimate of the proportion of adaptive mutations. As expected from the fact that more slightly deleterious mutations can be fixed in smaller populations, fewer adaptive mutations are inferred with smaller ancestral  $N_e$ -values (EYRE-WALKER 2002). Again, the mutation rate and mutational bias do not greatly affect the estimates. The combination of Figure 4 with the wide confidence intervals for  $P_a$  from Tables 1–3 raises the question of whether this approach can determine the presumably small fraction of adaptive mutations with any precision. Larger data sets and the use of the same sets of genes in the two species being compared may help to narrow the error bounds on the estimates.

However, potential difficulties still remain. One is the assumption of free recombination among variants. Linkage increases the rate of fixation of deleterious mutations while decreasing that for advantageous mutations (BIRKY and WALSH 1988); close linkage can have important effects even with weak selection (MCVEAN and CHARLESWORTH 2000; KIM 2004). But this assumption seems reasonable for *D. pseudoobscura* and its relatives, with their high rates of recombination and lack of linkage disequilibrium within genes (DOBZHANSKY and POWELL 1975; SCHAEFFER and MILLER 1993; YI *et al.* 2003). In addition, departures from equilibrium due to demographic effects may introduce errors into the estimates. As discussed in MATERIALS AND METHODS, our choice of  $\theta_W$  instead of  $\pi$  as a synonymous diversity estimator is intended to minimize the effects of the population expansion that seems to have occurred in

*D. pseudoobscura* (MACHADO *et al.* 2002; SCHAEFFER 2002). However, we cannot exclude population bottlenecks in the distant past that could have led to a higher contribution of deleterious mutations to  $K_A/K_S$  relative to  $\pi_A/\pi_S$  and thus have elevated the estimate of the proportion of adaptive substitutions (EYRE-WALKER 2002), a problem common to all methods used to date.

**The reliability of estimates of parameters of the distribution of mutational effects:** Methods of estimating selection parameters for deleterious mutations that assume a distribution of selection coefficients make inferences about the distribution for new mutations prior to the action of selection, on the basis of the properties of mutations that are segregating in populations; these have been exposed to a long history of selection. The arithmetic mean  $s$  for segregating mutations is obtained by summing the products of the selection coefficient at each site  $i$  by the corresponding frequency of heterozygotes  $\pi_{A_i}$  and normalizing by the summed frequencies of heterozygotes ( $\bar{s}_{\text{seg}} = \sum s_i \pi_{A_i} / \sum \pi_{A_i}$ ). For strong selection, Equation 2a implies that this is equal to the harmonic mean of the distribution of  $s$ -values for new mutations (ORR and KIM 1998; SUNYAEV *et al.* 2001).

Tables 2 and 3 show that the mean of the prior gamma distribution can be much larger than the harmonic mean for mutations above the effective neutrality threshold, indicating that much of the probability mass of the gamma distribution is far from the  $s$ -values representative of segregating mutations. This raises a serious issue concerning the meaning of inferences concerning the parameters of the gamma distribution; these are based on the properties of mutations that have little relation to the bulk of the mutations in the prior distribution.

With this caveat in mind, one of the strongest results from our arbitrary purifying selection model is not the exact set of parameter values themselves, but rather the exclusion of the large number of parameter combinations that are not compatible with the data. Our difficulties fitting distributions of mutational effects for  $c_n \geq 5\%$ , for example, probably suggest that  $<5\%$  of all nonsynonymous mutations stem from a small set of mutational effects distinct from the continuous distribution, which behave as neutral. Similarly, Tables 2 and 3 allow us to restrict the credible range for the shape parameter of a gamma distribution to shapes  $>0.1$  and usually  $<1$ , where 1 is equivalent to an exponential distribution; not many credible estimates have less leptokurtic shapes. This agrees with results for mitochondrial genes, based on a different approach (PIGANEAU and EYRE-WALKER 2003).

If our estimates of the arithmetic and harmonic means of the distribution of  $s$  are even approximately correct (of the order of  $10^{-4}$  and  $10^{-5}$ , respectively; see RESULTS), they imply that most deleterious nucleotide substitutions affecting protein sequences in our two species are subject to very weak selection. This seems

inconsistent with the classical estimates of harmonic mean heterozygous selection coefficients of the order of 1% in *D. melanogaster*, obtained by comparing inbreeding loads for viability with the mutational decline in mean viability, as well as with estimates of mean homozygous selection coefficients of the order of  $\geq 10\%$  obtained from mutation-accumulation experiments (CROW 1993; CHARLESWORTH and HUGHES 2000; CHARLESWORTH *et al.* 2004). The former are, however, biased upward by being weighted by the selection coefficients themselves, and the latter are known to be biased upward when there is a wide distribution of homozygous selection coefficients (CROW 1993). In addition, it is likely that insertional mutations that effectively knock out gene function, like those caused by transposable elements, contribute substantially to these estimates (KEIGHTLEY and EYRE-WALKER 1999; CHARLESWORTH *et al.* 2004). In contrast, the estimates of heterozygous selection coefficients of the order of  $10^{-3}$ , obtained from a population screen for null alleles at enzyme loci in *D. melanogaster* by LANGLEY *et al.* (1981), are consistent with our estimates, assuming that they represent the effects of loss of function at nonvital loci. As pointed out to us by Allen Orr (A. ORR, personal communication), it is difficult to reconcile the results of the null allele screen with the classical estimates of average selection coefficients for deleterious mutations.

**Perspectives for the future:** An obvious way to narrow the confidence intervals is the compilation of data sets with more genes. This does not, however, solve the conceptual problem that more degrees of freedom are needed to estimate more parameters. One way of obtaining additional degrees of freedom is to use shared polymorphisms to obtain an estimate of ancestral  $N_e$  (WAKELEY and HEY 1997). Since  $\sim 3\%$  of the polymorphisms in *D. miranda* and *D. pseudoobscura* seem to be shared by both species (CHARLESWORTH *et al.* 2005), this is in principle possible. In the present study, we have ignored this approach, due to limited statistical power (only three suitable loci have been surveyed in both species). Another possibility for estimating additional parameters is to use sets of three species where significant differences can be observed in  $K_A/K_S$  as well as in  $\pi_A/\pi_S$ . This might eventually allow simultaneous estimation of  $P_a$  and  $c_n$  as well as of the shape and location parameters of the distribution of mutational effects. We also deliberately did not classify amino acid changes according to conservative, radical, etc., to keep the model simple. There is no reason why this could not be done with larger data sets, by dividing observations of  $\pi_A$  into several classes according to an *a priori* classification of their effects on protein function (SUNYAEV *et al.* 2001; WILLIAMSON *et al.* 2005).

We thank Deborah Charlesworth, Stephen Schaeffer, and two anonymous reviewers for comments on the manuscript. This work was supported by grants from the Biotechnology and Biological Sciences Research Council, the Leverhulme Trust, and the Royal Society.

## LITERATURE CITED

- AKASHI, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**: 221–238.
- AKASHI, H., and S. W. SCHAEFFER, 1997 Natural selection and the frequency distributions of “silent” DNA polymorphism in *Drosophila*. *Genetics* **146**: 295–307.
- ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**: 257–268.
- BARTOLOMÉ, C., X. MASIDE, S. YI, A. L. GRANT and B. CHARLESWORTH, 2005 Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*. *Genetics* **169**: 1495–1507.
- BIERNE, N., and A. EYRE-WALKER, 2004 The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* **21**: 1350–1360.
- BIRKY, JR., C. W., and J. B. WALSH, 1988 Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**: 6414–6418.
- BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–908.
- BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. *Nature* **416**: 531–534.
- BUSTAMANTE, C. D., R. NIELSEN and D. L. HARTL, 2003 Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* **63**: 91–103.
- BUSTAMANTE, C. D., A. FLEDEL-ALON, S. WILLIAMSON, R. NIELSEN, M. T. HUBISZ *et al.*, 2005 Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.
- CHARLESWORTH, B., and K. A. HUGHES, 2000 The maintenance of genetic variation in life-history traits, pp. 369–392 in *Evolutionary Genetics: From Molecules to Morphology*, edited by R. S. SINGH and C. B. KRIMBAS. Cambridge University Press, Cambridge, UK.
- CHARLESWORTH, B., H. BORTHWICK, C. BARTOLOMÉ and P. PIGNATELLI, 2004 Estimates of the genomic mutation rate for detrimental alleles in *Drosophila melanogaster*. *Genetics* **167**: 815–826.
- CHARLESWORTH, B., C. BARTOLOMÉ and V. NOËL, 2005 The detection of shared and ancestral polymorphisms. *Genet. Res.* **86**: 149–157.
- CROW, J. F., 1993 Mutation, mean fitness, and genetic load, pp. 3–42 in *Oxford Surveys in Evolutionary Biology*, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.
- DOBZHANSKY, T., and J. R. POWELL, 1975 *Drosophila pseudoobscura* and its American relatives, *Drosophila persimilis* and *Drosophila miranda*, pp. 537–587 in *Handbook of Genetics: Invertebrates of Genetic Interest*, edited by R. C. KING. Plenum Press, New York.
- DRAKE, J. W., B. CHARLESWORTH, D. CHARLESWORTH and J. F. CROW, 1998 Rates of spontaneous mutation. *Genetics* **148**: 1667–1686.
- EYRE-WALKER, A., 2002 Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**: 2017–2024.
- FAY, J. C., G. J. WYCKOFF and C. I. WU, 2001 Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.
- HALDANE, J. B. S., 1924 The mathematical theory of natural and artificial selection. Part I. *Trans. Camb. Philos. Soc.* **23**: 19–41.
- HALDANE, J. B. S., 1927 The mathematical theory of natural and artificial selection. Part V: selection and mutation. *Proc. Camb. Philos. Soc.* **23**: 838–844.
- HARTL, D. L., E. N. MORIYAMA and S. A. SAWYER, 1994 Selection intensity for codon bias. *Genetics* **138**: 227–234.
- IHAKA, R., and R. GENTLEMAN, 1996 R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**: 299–314 (<http://www.r-project.org/>).
- KEIGHTLEY, P. D., and A. EYRE-WALKER, 1999 Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**: 515–523.
- KIM, Y., 2004 Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. *Mol. Biol. Evol.* **21**: 286–294.
- KIMURA, M., 1962 On the probability of fixation of mutant genes in a population. *Genetics* **47**: 713–719.
- KIMURA, M., 1971 Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.* **2**: 174–208.
- KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

- KIMURA, M., and T. OHTA, 1969a The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* **63**: 701–709.
- KIMURA, M., and T. OHTA, 1969b The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**: 763–771.
- LANGLEY, C. H., R. A. VOELKER, A. J. BROWN, S. OHNISHI, B. DICKSON *et al.*, 1981 Null allele frequencies at allozyme loci in natural populations of *Drosophila melanogaster*. *Genetics* **99**: 151–156.
- LI, W. H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**: 337–345.
- LI, W. H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- MACHADO, C. A., R. M. KLIMAN, J. A. MARKERT and J. HEY, 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* **19**: 472–488.
- MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**: 652–654.
- MCVEAN, G. A. T., and B. CHARLESWORTH, 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74**: 145–158.
- MCVEAN, G. A. T., and B. CHARLESWORTH, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**: 929–944.
- MCVEAN, G. A. T., and J. VIEIRA, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**: 245–257.
- NACHMAN, M. W., 1998 Deleterious mutations in animal mitochondrial DNA. *Genetica* **102/103**: 61–69.
- NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- NELDER, J. A., and R. MEAD, 1965 A simplex algorithm for function minimization. *Comput. J.* **7**: 308–313.
- ORR, H. A., and Y. KIM, 1998 An adaptive hypothesis for the evolution of the Y chromosome. *Genetics* **150**: 1693–1698.
- PIGANEAU, G., and A. EYRE-WALKER, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proc. Natl. Acad. Sci. USA* **100**: 10335–10340.
- POWELL, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York.
- PRZEWORSKI, M., J. D. WALL and P. ANDOLFATTO, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* **18**: 291–298.
- RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* **13**: 735–748.
- RAND, D. M., and L. M. KANN, 1998 Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* **102/103**: 393–407.
- ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP. DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
- R-PROJECT, 2005 *R: A Language for Data Analysis and Graphics*. (<http://www.r-project.org/>).
- SAWYER, S. A., 1994 Inferring selection and mutation from DNA sequences: the McDonald-Kreitman test revisited, pp. 77–87 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. GOLDING. Chapman & Hall, New York.
- SAWYER, S. A., and D. L. HARTL, 1992 Population genetics of polymorphism and divergence. *Genetics* **132**: 1161–1176.
- SAWYER, S. A., D. E. DYKHUIZEN and D. L. HARTL, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. USA* **84**: 6225–6228.
- SAWYER, S. A., R. J. KULATHINAL, C. D. BUSTAMANTE and D. L. HARTL, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57**: S154–S164.
- SCHAEFFER, S. W., 2002 Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. *Genet. Res.* **80**: 163–175.
- SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **135**: 541–552.
- SUNYAEV, S., V. RAMENSKY, I. KOCH, W. LATHE, III, A. S. KONDRASHOV *et al.*, 2001 Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**: 591–597.
- SUNYAEV, S. R., W. C. LATHE, III, V. E. RAMENSKY and P. BORK, 2000 SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16**: 335–337.
- TAJIMA, F., 1989 The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.
- WAKELEY, J., and J. HEY, 1997 Estimating ancestral population parameters. *Genetics* **145**: 847–855.
- WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–276.
- WEINREICH, D. M., and D. M. RAND, 2000 Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics* **156**: 385–399.
- WILLIAMSON, S., A. FLEDEL-ALON and C. D. BUSTAMANTE, 2004 Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **168**: 463–475.
- WILLIAMSON, S. H., R. HERNANDEZ, A. FLEDEL-ALON, L. ZHU, R. NIELSEN *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **102**: 7882–7887.
- WRIGHT, A., B. CHARLESWORTH, I. RUDAN, A. CAROTHERS and H. CAMPBELL, 2003 A polygenic basis for late-onset disease. *Trends Genet.* **19**: 97–106.
- YI, S., D. BACHTROG and B. CHARLESWORTH, 2003 A survey of chromosomal and nucleotide sequence variation in *Drosophila miranda*. *Genetics* **164**: 1369–1381.

Communicating editor: S. W. SCHAEFFER

#### APPENDIX: EXTENDING THE STRONG PURIFYING SELECTION MODEL

We can extend the strong purifying selection model as follows, to lighten the assumptions involved. The aim is to place bounds on the estimates of the fraction of nearly neutral mutations and the harmonic mean of  $s$ , for the species with the lower effective population size, *i.e.*, species 1. We use the approximate formula for the equilibrium diversity at sites under selection given by Equation 15 of MCVEAN and CHARLESWORTH (1999). A simple extension to this, using the formulation that led to Equation 5a, yields the following expression for species 1,

$$\frac{\pi_{A1}}{\pi_{S1}} \approx c_{nc} + (1 - c_{nc}) \int_{s>s'} \left( \frac{2}{\gamma_1} \right) \left\{ \frac{(e^{\gamma_1} - 1) + (1/2)(\kappa - 1)\gamma_1}{(\kappa + e^{\gamma_1})} \right\} \psi(s) ds, \quad (A1)$$

where  $\gamma_1 = 4N_{e1}s$  is the scaled measure of selection intensity for species 1,  $\psi(s)$  is the probability density of  $s$ , conditional on  $s$  falling outside the domain of effective neutrality (defined by the relation  $\gamma_1 \leq 4N_{e1}s' = 2$ ), and  $c_{nc}$  is the fraction of neutral and effectively neutral mutations for species 1 (see Table 2). Numerical integrations have

shown that Equation A1 typically predicts diversity patterns with a relative error of  $\sim 5\text{--}10\%$  when compared to our more accurate method.

A similar relation can be written for species 2 (the species with larger  $N_e$ ), retaining the same values of  $s'$  and  $\psi(s)$ . The only change is that  $N_{e_2}$  is substituted for  $N_{e_1}$  when specifying  $\gamma$  inside the integral, and  $c_{ne}$  in the first term on the right-hand side of the equation is replaced by  $c_{ne}\alpha$ , where  $\alpha < 1$ . The  $\alpha$  parameter reflects the fact that a larger fraction of mutations with  $s < s'$  do not behave as effectively neutral in species 2, so that  $s'$  does not constitute the border of effective neutrality in this species.

For our purposes, the term in braces is just a nuisance parameter, since we are interested only in  $N_{e_1s_h}$ , the harmonic mean of selection coefficients  $s > s'$  for species 1. Using the mean-value theorem, we can replace the integrals for the two species by

$$\frac{1}{(2N_{e_1s_h})} \frac{(e^{\bar{\gamma}_i} - 1) + (1/2)(\kappa - 1)\bar{\gamma}_i}{(\kappa + e^{\bar{\gamma}_i})} = \frac{I_i}{(2N_{e_1s_h})},$$

where  $\bar{\gamma}_i$  is a value of  $\gamma_i$  in the domain of integration above  $s'$ , and  $s_h$  is the harmonic mean of  $s$  with respect to  $\psi(s)$  over this domain, *i.e.*, the harmonic mean of  $s$  for amino acid mutations other than effectively neutral ones in our focal species, species 1.

$I_i$  is an increasing function of  $\bar{\gamma}_i$  for  $\kappa \geq 0$  and  $\bar{\gamma}_i \geq 0$ , so that a lower bound is obtained by setting  $\bar{\gamma}_i$  to  $\gamma_i' = 4N_{e_1}s'$ . The upper bound is 1; in the present case, this is very close to the actual value of  $I_2$  and is used in its place. By the same argument that led to Equations 2a and 3a, and using the lower bound of  $I_1$  and the upper bound of  $I_2$ , together with the fact that  $\alpha < 1$ , after some algebra we obtain a lower bound to the estimate of  $c_{ne}$ ,

$$\hat{c}_{ne} = \frac{2(I_1 - 1)/(\gamma_1'(r - 1)) + \hat{c}_n}{\{1 + 2(I_1 - 1)/(\gamma_1'(r - 1))\}} \quad (\text{A2})$$

where  $\hat{c}_n$  is the estimate of  $c_n$  from Equation 3a, and  $r$  is the ratio of silent diversity for species 2 to that for species 1. This expression can in turn be used to yield a lower-bound estimate for  $N_{e_1s_h}$  by using Equations 1a and 1b:

$$2N_{e_1s_h} \geq \frac{I_1(1 - \hat{c}_{ne})}{\{(\pi_{A1}/\pi_{S1}) - \hat{c}_{ne}\}} \quad (\text{A3})$$

Similarly, approximate estimates of  $c_a$  and  $P_a$  are obtained by substituting (A2) into Equations 4.

This approach provides a conservative method for improving on the assumption of strong purifying selection, without having to make specific assumptions about the distribution of mutational effects. Application of this method to the weighted data on *D. miranda* and *D. pseudoobscura*, assuming  $\gamma_1' = 2$  and a mutational bias of 2, gave estimates of  $c_{ne}$  of 4.5% ( $-1.6\%/11.6\%$ ),  $N_{e_1s_h}$  of 3.00 (1.76/16.7), and an estimate of  $P_a$  of 51% ( $-60\%/117\%$ ) for *D. miranda* (the terms in parentheses give the approximate lower and upper bootstrap fifth percentiles). The assumptions used to derive (A2) mean that this approach cannot be used for the species with the larger  $N_e$ .

Use of Equation 3a of the text provides an upper-bound estimate of  $c_{ne}$ , since it assumes that all deleterious mutations outside the effectively neutral range have  $s$ -values sufficiently large that the deterministic expression in Equation 2a is valid, ignoring mutations with very small  $s$ -values.  $1/s_h$  in Equation 2a must therefore be smaller than the values allowing for a wider distribution of  $s$ -values, and so  $c_n$  must be larger. The true value of  $c_{ne}$  is thus likely to lie between the estimates from Equations 3a and A2.

Given the uncertainties involved, the degree of concordance between the estimates of  $N_{e_1s_h}$  from this approximate method and those in Tables 1–3 is encouraging, supporting the conclusion that there must be sufficiently strong selection against deleterious amino acid substitutions for the harmonic mean of  $N_e s$  to be substantially  $> 1$ , even in *D. miranda* with its relatively low effective population size. Unfortunately, the percentile intervals for  $P_a$  are so wide that no confidence can be placed in the relevant estimates.

An even more conservative lower bound on  $N_{e_1s_h}$  for nonsynonymous mutations above the threshold of effective neutrality is given by setting  $c_{ne}$  to zero in expression (A3); this has the advantage of using polymorphism data on only one species, which makes it widely applicable. With a neutrality threshold of  $\gamma_1 = 2$  and with  $\kappa = 2$ ,  $I_1 = 0.787$ , so we get  $N_{e_1s_h} \geq 2.15$  ( $N_{e_2s_h} \geq 12.5$ ) in the present case.

This estimate can be applied to any suitable data set on coding sequence polymorphisms, as long as a credible assumption about  $c_{ne}$  can be made. For example, SUNYAEV *et al.* (2000) reported an estimate of 0.33 for the ratio of diversity at nondegenerate coding sites to fourfold degenerate sites, in a large-scale survey of EST-based human SNPs. With  $\kappa = 2$  and  $I = 0.787$ , we get  $N_{e_1s_h} \geq 1.18$  for  $c_{ne} = 0$ . With an effective population size for humans of  $\sim 10,000$ , this suggests a harmonic mean selection coefficient for nonneutral amino acid variants of at least  $1.18 \times 10^{-4}$ . Other studies have suggested that  $\sim 20\%$  of amino acid variants in humans are effectively neutral (FAY *et al.* 2001; SUNYAEV *et al.* 2001); if this value of  $c_{ne}$  is used in expression (A3), the estimate of  $N_{e_1s_h}$  for nonneutral variants becomes 2.42.

### Estimating selection on non-synonymous mutations

Laurence Loewe, Brian Charlesworth, Carolina Bartolomé and Véronique Noël

Genetics (Accepted Nov 2005)

### Supplementary Online Material

#### Genes used in the study:

*D. miranda*: *ade3*<sup>1</sup>, *Adh*<sup>2</sup>, *amd*<sup>3</sup>, *bcd*<sup>4</sup>, *Bruce*<sup>1</sup>, *Ddc*<sup>1</sup>, *Eno*<sup>1</sup>, *Gapdh2*<sup>1</sup>, *Gld*<sup>1</sup>, *hyd*<sup>1</sup>,  
*Lam*<sup>1</sup>, *rosy*<sup>1</sup>, *scute*<sup>2</sup>, *sesB*<sup>2</sup>, *sisA*<sup>2</sup>, *sry-alpha*<sup>2</sup>, *Uro*<sup>1</sup>

1. Bartolomé, C., X. Maside, S. Yi, A. L. Grant and B. Charlesworth, 2005 Patterns of selection on synonymous and non-synonymous variants in *Drosophila miranda*. *Genetics* 169: 1495-1507.
2. Yi, S., D. Bachtrog and B. Charlesworth, 2003 A survey of chromosomal and nucleotide sequence variation in *Drosophila miranda*. *Genetics* 164: 1369-1381.

*D. pseudoobscura*: *Adh*<sup>1,2</sup>, *Amy1*<sup>3</sup>, *bcd*<sup>4</sup>, *EcR*<sup>5</sup>, *Est-5A*<sup>6</sup>, *Est-5B*<sup>7</sup>, *Est-5C*<sup>6</sup>, *eve*<sup>5</sup>, *exu*<sup>1,2</sup>, *Hsp82*<sup>8</sup>, *per*<sup>9</sup>, *rh1*<sup>4</sup>, *run*<sup>10</sup>, *rosy*<sup>11</sup> (Current names for the genes *Hsp82* and *rh1* are *Hsp83* and *ninaE*, respectively.)

1. Schaeffer, S. W., and E. L. Miller, 1992 Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* 132: 163-178.
2. Schaeffer, S. W., 2002 Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. *Genet. Res.* 80: 163-175.

3. Popadic, A., and W. W. Anderson, 1994 The history of a genetic system. *Proc. Natl. Acad. Sci. USA* 91: 6819-6823.
4. Machado, C. A., R. M. Kliman, J. A. Markert and J. Hey, 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* 19:472-88.
5. Schaeffer, S.W., M. P. Goetting-Minesky, M. Kovacevic, J. R. Peoples, J. L. Graybill, J. M. Miller, K. Kim, J. G. Nelson, W. W. Anderson, 2003 Evolutionary genomics of inversions in *Drosophila pseudoobscura*: evidence for epistasis. *Proc. Natl. Acad. Sci. USA* 100: 8319-8324.
6. King, L. M., The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. *Genetics* 148: 305-315.
7. Veuille, M, and L. M. King, 1995 Molecular basis of polymorphism at the esterase-5B locus in *Drosophila pseudoobscura*. *Genetics* 141: 255-262.
8. Wang R. L., J. R. Wakeley and J. Hey, 1997 *Genetics* 147: 1091-1161.
9. Wang R. L., and J. Hey, 1996 The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the *period* locus. *Genetics* 144: 1113-1126.
10. Kovacevic, M., S.W. Schaeffer, 2000 Molecular population genetics of X-linked genes in *Drosophila pseudoobscura*. *Genetics* 156: 155-172.
11. Begun, D. J., and P. Whitley, 2002 Molecular population genetics of *Xdh* and the evolution of base composition in *Drosophila*. *Genetics* 162: 1725-1735.
12. Luk, S. K., Kilpatrick, M., Kerr, K. & Macdonald, P. M., 1994 Components acting in localization of *bicoid* mRNA are conserved among *Drosophila* species. *Genetics* 137: 521-30.

**Supplementary Table 1:** Summary of the data for *D. pseudoobscura*.

Columns denote:

- A:** Gene name, total number of sequences analyzed, number of codons, total number of synonymous base pairs, number of polymorphic synonymous base pairs, total number of non-synonymous base pairs, number of polymorphic non-synonymous base pairs, total number of silent sites, number of polymorphic silent sites, chromosome where the gene is located.
- B:** Average pairwise diversities (Pi) at synonymous sites (S), non-synonymous sites (A) and silent sites (s) are given together with the ratios of uncorrected Pi(A)/Pi(S).
- C:** WATTERSON'S estimator  $\theta_w$  (Theta). Rest, see **B**.
- StDev gives the standard deviation and StErrMean the standard error of the arithmetic mean.

<b>A</b>	No of sqces	No codons analyzed	No of S sites analyzed	N(S)	No of A sites analyzed	N(A)	No of s sites analysed	N(s)	Chr
Adh/Adh-r	139	529	385	132	1202	16	1948	439	4
Amy1	7	494	350	24	1132	9	825	52	3
bcd	21	379	278	33	859	12	475	45	2
EcR	103	87	63	10	198	0	132	21	3
Est-5A	8	546	390	24	1248	19	1283	49	2
Est-5B	17	545	385	69	1250	34	909	88	2
Est-5C	8	545	382	33	1253	13	1066	53	2
eve	101	101	76	14	227	1	147	24	3
exu-1	106	77	52	3	179	0	176	14	3
Hsp82	12	240	158	6	565	1	1305	34	X
per	16	314	228	26	735	7	548	53	XL
rh1	18	312	220	16	716	2	725	85	2
run	40	41	29	5	94	0	316	48	XL
rosy	9	520	390	45	1170	16	390	45	2

<b>B</b>	Pi(S)	Pi(A)	Pi(A)/Pi(S)	Pi (s)
Adh/Adh-r	0.0269	0.0013	0.0480	0.0181
Amy1	0.0269	0.0033	0.1218	0.0261
bcd	0.0246	0.0024	0.0974	0.0186
EcR	0.0167	0.0000	0.0000	0.0173
Est-5A	0.0217	0.0043	0.1983	0.0144
Est-5B	0.0401	0.0068	0.1687	0.0211
Est-5C	0.0318	0.0037	0.1157	0.0180
eve	0.0249	0.0001	0.0036	0.0236
exu-1	0.0025	0.0000	0.0000	0.0074
Hsp82	0.0079	0.0005	0.0685	0.0059
per	0.0277	0.0021	0.0754	0.0227
rh1	0.0149	0.0003	0.0208	0.0189
run	0.0199	0.0000	0.0000	0.0168
rosy	0.0338	0.0041	0.1200	0.0338
Mean	0.0229	0.0021	0.0742	0.0188
Variance	0.000101	0.000004	0.004310	0.000050
StDev	0.0100	0.0021	0.0656	0.0070
StErrMean	0.0027	0.0006	0.0175	0.0019

<b>C</b>	Theta(S)	Theta(A)	Theta(A)/Theta(S)	Theta(s)
Adh/Adh-r	0.0622	0.0024	0.0386	0.0409
Amy1	0.0280	0.0033	0.1162	0.0257
bcd	0.0330	0.0039	0.1176	0.0263
EcR	0.0303	0.0000	0.0000	0.0305
Est-5A	0.0237	0.0059	0.2474	0.0147
Est-5B	0.0530	0.0081	0.1519	0.0286
Est-5C	0.0333	0.0040	0.1200	0.0192
eve	0.0357	0.0009	0.0238	0.0315
exu-1	0.0111	0.0000	0.0000	0.0152
Hsp82	0.0126	0.0006	0.0468	0.0086
per	0.0309	0.0027	0.0868	0.0291
rh1	0.0211	0.0008	0.0384	0.0341
run	0.0399	0.0000	0.0000	0.0357
rosy	0.0425	0.0050	0.1185	0.0425
Mean	0.0327	0.0027	0.0790	0.0273
Variance	0.000197	0.000006	0.005071	0.000099
StDev	0.0140	0.0025	0.0712	0.0099
StErrMean	0.0037	0.0007	0.0190	0.0027