Characterizing the Influence of Effective Population Size on the Rate of Adaptation: Gillespie's Darwin Domain

Jeffrey D. Jensen*,[1], and Doris Bachtrog[2]

[1]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2]Department of Integrative Biology, University of California, Berkeley

*Corresponding author: E-mail: jeffrey.jensen@epfl.ch.

Accepted: 17 June 2011

Abstract

Characterizing the role of effective population size in dictating the rate of adaptive evolution remains a major challenge in evolutionary biology. Depending on the underlying distribution of fitness effects of new mutations, populations of different sizes may differ vastly in their rate of adaptation. Here, we collect polymorphism data at over 100 loci for two closely related Drosophila species with different current effective population sizes (N_e), *Drosophila miranda* and *D. pseudoobscura*, to evaluate the prevalence of adaptive evolution versus genetic drift in molecular evolution. Utilizing these large and consistently sampled data sets, we obtain greatly improved estimates of the demographic histories of both species. Specifically, although current N_e differs between these species, their ancestral sizes were much more similar. We find that statistical approaches capturing recent adaptive evolution (using patterns of polymorphisms) detect higher rates of adaptive evolution in the larger *D. pseudoobscura* population. In contrast, methods aimed at detecting selection over longer time periods (i.e., those relying on divergence data) estimate more similar rates of adaptation between the two species. Thus, our results suggest an important role of effective population size in dictating rates of adaptation and highlight how complicated population histories—as is probably the case for most species—can effect rates of adaptation. Additionally, we also show how different methodologies to detect positive selection can reveal information about different timescales of adaptive evolution.

Key words: selective sweeps, demography, natural selection.

Introduction

Understanding the relative role of effective population size on the rate of adaptation has been of long standing interest to evolutionary biologists. Depending on the distribution of fitness effects (DFE) of new mutations, Gillespie defined three specific model-based domains of molecular evolution (Gillespie 1999, 2001). In the Ohta domain (Ohta 1973), patterns of molecular evolution are driven mainly by slightly deleterious mutations (Gillespie 1999). Under this model, the rate of substitution decreases with increasing effective population size, due to an increase in the efficiency of purifying selection against deleterious mutations. In the Kimura domain (Kimura 1968), molecular evolution is dominated by mutations with no effect on fitness, and the rate of substitution is independent of the effective population size but simply given by the neutral mutation rate (Gillespie 1999). Finally, in the Darwin domain, molecular evolution is driven by beneficial mutations, and the rate of substitution is predicted to increase with effective population size (Gillespie 1999). If beneficial mutations are independent, rates of adaptation increase linearly with increasing population size. However, if beneficial mutations are common and linked, the rate of substitution will be substantially reduced and eventually become independent of the effective population size of a species (Gillespie 2000). The relationship between population size and rates of molecular evolution is additionally complicated by the fact that positive selection may actually increase the rate of fixation of deleterious substitutions at linked sites (Bachtrog and Gordo 2004). Thus, depending on the underlying DFE and other population parameters, different patterns in rates of molecular evolution are expected with changing population size.

With the advent of large-scale genomics, a tremendous amount of both data and methodology has recently been published to address the underlying DFE of new mutations. In particular, a number of recent studies in Drosophila have

found that positive selection may be prevalent (Darwin domain) but have resulted in vastly different estimates of the underlying distribution of selection coefficients (see recent review of Sella et al. 2009; Sattath et al. 2011). Assuming that the observed correlation between recombination rate and nucleotide diversity in Drosophila is driven by beneficial mutations, Eyre-Walker (2006) estimates the joint parameter, $350 < N_e s < 3{,}500$. Macpherson et al. (2007) fit levels of neutral polymorphism to divergence and concluded that selection is stronger, with $N_e s \sim 10^5$. Using likelihood based and approximate Bayesian based methodologies, respectively, Li and Stephan (2006) and Jensen, Thornton, and Andolfatto (2008) estimate $N_e s \sim 1{,}000$. Hence, based upon these estimates, most Drosophila populations are expected to fall well within the Darwin domain, where the rate of substitution is positively correlated with effective population size. However, a consensus is far from being reached. A number of recent studies, all using McDonald–Kreitman (MK)-like statistical procedures to infer selection (i.e., comparing synonymous and nonsynonymous polymorphisms with divergence; McDonald and Kreitman 1991), estimated $N_e s < 100$ in Drosophila (Sawyer et al. 2003; Andolfatto 2007; Eyre-Walker and Keightley 2009).

Broadscale biological comparisons corroborate at least some correlation between the rate of adaptation and effective population size—hominids appear to be undergoing relatively little adaptive evolution, Drosophila and rodent species with their generally larger population sizes are estimated to show intermediate levels of adaptation, whereas bacteria with their very large populations display high rates of adaptive evolution (e.g., Kimura 1983; Nielsen and Yang 2003; Bustamante et al. 2005; Nielsen et al. 2005; Charlesworth and Eyre-Walker 2006; Halligan et al. 2009). However, other species such as yeast and Arabidopsis, which have population sizes similar to Drosophila (or larger), show little evidence of adaptive amino acid evolution, though differences in mating systems may be confounding these comparisons (Bustamante et al. 2002).

In order to avoid confounding effects of vastly different biological systems with very different life histories, it is desirable to compare species at a much closer phylogenetic scale. Bachtrog (2008) examined two Drosophila species—*Drosophila miranda* and *D. melanogaster*—that show a roughly 5-fold difference in their levels of neutral diversity, to evaluate the influence of effective population size on rates of adaptive evolution. Analyzing nearly 100 nonhomologous X-linked loci in both species, Bachtrog estimated a similar fraction of amino acid mutations being driven to fixation by selection between the two species. Thus, more consistent with estimates of strong selection, this analysis suggests that effective population size may not be a major determinant in rates of protein adaptation.

However, there are a number of possible complications with this conclusion. Although levels of neutral diversity are lower in *D. miranda* relative to *D. melanogaster*, this measure only accounts for recent effective population sizes, and it is plausible that the historical population size of both species may indeed have been more similar. This is consistent with evidence of a recent population size reduction in *D. miranda* (Yi et al. 2003; Bachtrog and Andolfatto 2006; Bachtrog 2008). Additionally, estimation was performed using a divergence-based procedure with *D. pseudoobscura* as an outgroup. Thus, much of the adaptive evolution inferred in *D. miranda* may actually reflect selection in the much larger population of *D. pseudoobscura* because divergence also includes fixations along the *D. pseudoobscura* lineage. Further, although both *D. melanogaster* and *D. miranda* belong to the genus Drosophila, they are fairly diverged at the DNA sequence level, live in different environments, and may have very different life history strategies. Finally, the loci compared in the two species represented a nonoverlapping data set consisting of nonhomologous loci.

In order to revisit the debate on the influence of effective population size on rates of molecular evolution, we gathered polymorphism data for over 100 X-linked loci in two closely related species with different effective population sizes, *D. miranda* and *D. pseudoobscura* (e.g., as recently demonstrated by Loewe et al. 2006, in the context of characterizing the relative strength of purifying selection between the two species). This comparison circumvents the problems discussed above and presents several advantages: 1) *D. miranda* and *D. pseudoobscura* appear to have a greater difference in their effective population sizes—with estimates suggesting a difference of almost an order of magnitude (Loewe et al. 2006), which should increase our chance to detect the influence of N_e; 2) *D. miranda* and *D. pseudoobscura* are two closely related sister species that are morphologically indistinguishable and diverged only about 2 Mya, suggesting that they may share more similar life histories; 3) we employ a consistently collected data set consisting of over 100 homologous genes sampled in both species which ensures that we compare genes that evolve under similar constraints in both species; 4) we explicitly model the demographic history of both species using a recently proposed likelihood-based demographic estimator and—rather than relying on standard equilibrium–based models—use this inferred demographic history to inform our inference of selection operating in both species; 5) in addition to divergence-based approaches to quantify positive selection, we also employ a recently proposed polymorphism-based method to independently estimate the rate and strength of adaptive evolution, which also allows us to estimate parameters of adaptive evolution as distributions rather than fixed values. Utilizing our improved experimental design and methods, we find a significant difference in rates of adaptive evolution between these two species, with *D. pseudoobscura* evolving at a greater rate, at least in its

recent history. These results are discussed with regard to changes in population size, emphasizing the difference in the short- and long-term effective population size, and its influence on different methodologies commonly employed for detecting selection.

Methods

Survey of Coding Regions and Data Processing

Here, we present polymorphism data for 112 gene fragments in *D. miranda* and 123 gene fragments in *D. pseudoobscura*. Almost all genes are orthologous between the two species and are located on the X chromosome and were selected randomly with regards to function. Average sample size was 14 individuals in both species, and the average length surveyed for each locus was roughly 1 kb. The screen in *D. pseudoobscura* was conducted in order to match that published in *D. miranda* by Bachtrog et al. (2009). Details of polymerase chain reaction (PCR) primers are available from the authors upon request. Information about the individual loci surveyed and the geographic origin of the *D. miranda* strains investigated can be found in Bachtrog et al. (2009). The *D. pseudoobscura* population investigated was from Mesa Verde, Colorado, and was kindly provided by A. Larracuente.

Standard PCR procedures were used to amplify each region from genomic DNA from single male flies. PCR products were cleaned using Exonuclease I and Shrimp Alkaline Phosphatase and sequenced on both strands with the original PCR primers and internal sequencing primers if necessary, using Big-Dye (Version 3, Applied Biosystems). Sequence reactions were cleaned with sephadex plates (Edge Biosystems) and run on an ABI 3730 capillary sequencer. Chromatograms were edited and assembled using Sequencher (Gene Codes) software, and multiple sequence alignments were generated using MUSCLE (http://www.drive5.com/muscle/) with protein-alignment–assisted adjustments to preserve reading frames. Exon–intron boundaries were determined from the *D. pseudoobscura* genome sequence annotation (release 2.0). The sequences can be found under Genbank accession numbers (FN252903-FN256223).

A library of Perl scripts were used to calculate the estimated number of synonymous sites, average pairwise diversity (π) and average pairwise divergence (K) to the outgroup species (either *D. pseudoobscura* or *D. miranda*). A Jukes–Cantor correction was used to correct π and K for multiple hits. To infer lineage-specific divergence, we reconstructed a *D. miranda–D. pseudoobscura* ancestor (ANC) sequence using the maximum-likelihood approach implemented in the "codeml" program of PAML (Yang 1997). We either used *D. affinis* sequence (see Bachtrog 2008) or *D. athabasca* sequence (provided by K. Wong) as a more distant outgroup sequence. We were able to reconstruct the ANC for 107 *D. miranda* polymorphism loci and 119 *D. pseudoobscura* loci. Insertion–deletion polymorphisms and polymorphic sites overlapping alignment gaps were excluded from the analysis. Note that we make no distinction between ancestral and derived polymorphisms (i.e., mutations segregating in the ancestral population of *D. miranda* and *D. pseudoobscura* vs. newly arising ones after their split). Although ancestral polymorphism may be important when comparing such closely related species—extending perhaps 10 N_e generations since the split time (Clark 1997; Charlesworth et al. 2005)—recent analysis, however, suggests that ancestral polymorphisms represent only a small fraction of observed variation in *D. pseudoobscura* and *D. miranda* (Charlesworth et al. 2005). Also, ancestral polymorphism would make this species pair look more similar to each other with regards to polymorphism-based inferences and thus be conservative with regards to our conclusions (see below). Scripts used for data processing are available for download at: http://ib.berkeley.edu/labs/bachtrog/data/polyMORPHOrama/polyMORPHOrama.html.

Estimating Demographic Models

Using a recently proposed likelihood-based estimator (Gutenkunst et al. 2009) and taking advantage of our large consistently sampled polymorphism data sets in two closely related species of *Drosophila*, we estimate demographic models for *D. miranda* and *D. pseudoobscura*. These models are used as a baseline to calculate relevant critical values when inferring selection in these two species. By letting θ correspond to the parameters of a demographic model that one wishes to estimate from the observed frequency spectrum (denoted as $S[d_i, d_j, \ldots]$) and assuming no linkage between polymorphisms, each entry is an independent Poisson variable, with mean $M[d_i, d_j, \ldots]$. A likelihood function is then constructed as:

$$L(\theta|S) = \prod_{i=0\ldots P} \prod_{d_i=0\ldots n_i} \frac{e^{-M[d_i,d_j,\ldots d_p]} M[d_i, d_j, \ldots, d_p]^{S[d_i,d_j,\ldots d_p]}}{S[d_i, d_j, \ldots, d_p]!}.$$

Thus, using a diffusion approach, the expected allele frequency spectrum M is calculated under a particular demographic model. The similarity between M and the observed spectrum, S, is maximized over the values of θ.

Full code and documentation to implement dadi are available at: http://code.google.com/p/dadi/.

To consider the impact of fluctuating population size on the rate of adaptation and to infer how different approaches to detect selection are sensitive to such fluctuations (and thus show a different dependency on the long-term vs. current population size), we utilize a forward simulation approach, and condition on the demographic parameters estimated with dadi (see Results). Additionally, these simulations can be used to quantify how the smaller coalescent effective population size of *D. miranda* decreases our power to detect selection.

Demographic simulations, incorporating distributions of selection coefficients, were performed using the simulation program SFScode (Hernandez 2008). Briefly, the program is a generalized Wright–Fisher forward simulation approach for models with selection, recombination, and demography. The demographic history for each species is modeled as estimated with dadi (table 2), and the rate and strength of selection are taken from the recurrent hitchhiking estimates obtained from the Jensen, Thornton, and Andolfatto (2008) approach (table 4). The program and documentation are available for download at: http://sfscode.sourceforge.net/SFS_CODE/SFS_CODE_home/SFS_CODE_home.html.

Patterns of diversity suggest differences in current N_e between *D. miranda* and *D. pseudoobscura*; however, the dadi estimator suggests a recent severe bottleneck for *D. miranda* (see Results) but a more similar long-term N_e for most of their history. An approach to estimate differences in the long-term effective population size between species is to utilize patterns of codon usage, as proposed by Bulmer (1991). For two alleles, B_1 and B_2, Bulmer supposes that an individual carrying B_2 has a relative fitness $1 - s$, such that s is the selective advantage of B_2 compared with B_1. Utilizing a classic result from population genetics (Wright 1931; Crow and Kimura 1970):

$$f(p) \propto e^{Sp} p^{V-1} (1-p)^{U-1},$$

where $S = 2N_e s$, $V = 2N_e v$, and $U = 2N_e u$ (where u = the mutation rate from B_1 to B_2 and v = the mutation rate from B_2 to B_1), Bulmer notes that for $U + V$ large, the distribution will be clustered at the deterministic equilibrium. If small, the population is likely at or near one of the boundaries. Thus, the expected gene frequency is the probability of being near 1 rather than 0:

$$P = e^S V / (e^S V + U)$$

Thus, in a large population, a polymorphism is expected at every codon position, with a fraction of (P) B_1 codons and ($1 - P$) B_2 codons. In a small population, a fraction P of the relevant positions are monomorphic for B_1 and ($1 - P$) for B_2. Using the above equation, we can thus relate codon usage with population size. Assuming that $u = v$ and $S = \ln P(1 - P)$, N_e may be estimated. For this analysis, all sampled synonymous sites were considered.

We compare our results with those of Bachtrog and Andolfatto (2006), who recently fit demographic models to *D. miranda* polymorphism data (growth and bottlenecks). Under the growth model, $N/N_0 = e^{rT}$, where N is the current population size, N_0 is the ancestral population size, r is the growth rate, and T is the time at which growth began. They estimate a 5-fold growth with growth rate = 10 starting $0.161N_e$ generations in the past. Under a bottleneck model, the ancestral population, N_0, is reduced to size N_b at time T for d generations, at which point it recovers to size N_0, where $N_b = fN_0$. They estimate $f = 0.001$, $T = 0.08N_e$ generations ago, and $d = 0.004N_e$ generations.

Fitting Single Hitchhiking Models

Several statistical tests to identify recent adaptive evolution were applied to genes from both species. The composite likelihood ratio test (CLRT) (Kim and Stephan 2002) uses the spatial distribution of mutation frequencies in a genomic region and levels of variability among a population sample of DNA sequences to test for evidence of a selective sweep. This method compares the ratio of the composite likelihood of the data under the standard neutral model of constant population size, neutral evolution, and random mating, L_N (Data), to the composite likelihood of the data under the model of a selective sweep, $L_S(\hat{\alpha}, \hat{X}|\text{Data})$, where α is the maximum likelihood estimate (MLE) of $2Ns$ (where N is the effective population size and s the selection coefficient) and X is the MLE of the location of the beneficial mutation. The CLRT statistic employed is $\Lambda_{KS} = \log \frac{L_S(\hat{\alpha}, \hat{X}|\text{Data})}{L_N(\text{Data})}$. The null distribution of Λ_{KS} is obtained for each region by applying the CLRT to data sets obtained from simulations under the standard neutral model (using the program ms, Hudson 2002) with the observed region length (L) and θ. The recombination rate ρ per site is set at 8.8×10^{-8} per site per generation (Bachtrog 2008). For each locus, 1,000 neutral replicates were simulated using locus-specific parameters in order to assess significance. A complete users manual, as well as all necessary code, can be found at: http://www.yuseobkim.net/YuseobPrograms.html. The neutral model is rejected at level γ (5% used here) when the observed Λ_{KS} is greater than the $100(1 - \gamma)$ percentile of the null distribution.

The CLRT is sensitive to deviations from the assumptions of the standard neutral model, with population substructure and recent bottlenecks leading to a high false-positive rate (Jensen et al. 2005). As one approach to examining the potential effects of demography, we considered the demographic models estimated here from dadi in order to calculate more realistic cutoff values for evaluating statistical significance. As a second approach to assess the fit of individual loci to a selective sweep model, we also employed a goodness-of-fit (GOF) test that contrasts the null hypothesis H_0 that the data are drawn from a selection model as simulated by the CLRT, to the alternative hypothesis H_A that the data are not drawn from such a model (Jensen et al. 2005). A composite-likelihood scheme is used to approximate the probability of the data given the null, $P(\text{Data} \mid H_0)$, to the probability of the data given the alternative, $P(\text{Data} \mid H_A)$, on the basis of the site-frequency spectrum of mutations. Simulations (using the program ssw, Kim and Stephan 2002) under the null hypothesis are used to find the critical value of the CLRT GOF statistic for each region, with locus-specific (maximum likelihood) estimates

of α and X. Note that in this instance, the null model is a selective sweep as this test is employed conditional on rejecting the CLRT (Jensen et al. 2005). The program is available for download at: http://www.yuseobkim.net/YuseobPrograms.html.

In addition to skewing the frequency spectrum, positive selection may also result in strong linkage disequilibrium (LD) flanking the target of selection and reduced LD across the target (Kim and Nielsen 2004; Stephan et al. 2006; Jensen et al. 2007). We thus employ patterns of LD to test for selection at individual loci using the ω_{max} test (Kim and Nielsen 2004). The ω-statistic, which is defined as

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1}\left(\sum_{i,j\in L} r_{ij}^2 + \sum_{i,j\in R} r_{ij}^2\right)}{(1/l(S-l))\sum_{i\in L, j\in R} r_{ij}^2},$$

divides the S polymorphic sites in the data set into two groups, one from the first to the lth polymorphic site from the left and the other from the $(l+1)$th to the last site ($l = 2, \ldots, S-2$), where L and R represent the left and right set of polymorphic sites, and r_{ij}^2 is the squared correlation coefficient between the ith and jth sites. Thus, ω increases with increasing LD within each of the two groups and decreasing LD between the two groups (i.e., the larger the value of the statistic the more "sweep-like" the underlying pattern). For a locus, the value of l that maximizes ω (ω_{max}) is found. Singletons were excluded prior to calculation. The null distribution of ω for each genomic region is obtained from simulation under the standard neutral model (using the program ms (Hudson 2002) with fixed θ and L). As above, we set $\rho = 8.8 \times 10^{-8}$ per site per generation. As with the CLRT statistic, it is also possible to construct the null considering the estimated demographic models of *D. miranda* and *D. pseudoobscura*. The program is available for download at: http://www.molpopgen.org/software/libsequence.html.

Fitting Models of Recurrent Hitchhiking

To estimate selection parameters under a recurrent hitchhiking model, we use the approximate Bayesian approach of Jensen, Thornton, and Andolfatto (2008) (and see Thornton 2009). The level of reduction in variation due to recurrent selection depends on the joint parameter $2Ns\lambda$ (Wiehe and Stephan 1993). Both the rate, $2N\lambda$, and the fitness effect, s, of recurrent selection are estimated based upon their relationship with the means and standard deviations of common polymorphism summary statistics (the mean average pairwise diversity (π), the number of segregating sites (S), θ_H, and ZnS (Jensen, Thornton, and Andolfatto 2008)). Calculating these summary statistics from the observed data and from simulated data with parameters drawn from uniform priors, we implement the regression approach of Beaumont et al. (2002), which fits a local linear regression of simulated parameter values to simulated summary statistics, and substitutes the observed statistics into a regression equation. The prior distributions used were $s \sim$ Uniform (1.0×10^{-6}, 1.0) and $4N_e\lambda \sim$ Uniform (1.0×10^{-7}, 1.0×10^{-1}), and the tolerance, $\epsilon = 0.001$. Estimation is based on 10^6 draws from the prior using the recurrent selective sweep coalescent simulation machinery described in Jensen, Thornton, and Aquadro (2008). Briefly, sweeps are occurring in the genome at a rate determined by $2N\lambda$, where λ is the rate of sweeps per generation. In the simulations, sweeps are allowed both within the sampled region (of size M) as well as at linked sites. The rate of sweeps within a region is thus $2N\lambda M$, and each sweep may affect up to $4Ns/\rho_{bp}$. We set $\rho = 8.8 \times 10^{-8}$ per site per generation. For inferences on selection parameters, we assume exponential distributions of $2N\lambda$ and s, such that each draw from the prior represents the mean of the distribution. A complete users manual, as well as all necessary code, can be found at: http://www.molpopgen.org/software/JensenThorntonAndolfatto2008/.

Polymorphism- and Divergence-Based Methods to Infer Selection

To compare polymorphism and divergence, we implemented the MK test (McDonald and Kreitman 1991). Briefly, this approach considers a 2×2 contingency table of polymorphic synonymous and nonsynonymous variation, with synonymous and nonsynonymous divergence. With the sequence polymorphism data for both *D. miranda* and *D. pseudoobscura*, it is possible to consider true fixed differences, avoiding issues of estimating divergence based on a single sample. Additionally, the reconstructed ANC sequence allows us to estimate lineage-specific selection. P values are calculated using a Fisher's exact test.

We also apply a multilocus maximum likelihood version of the HKA test (Hudson et al. 1987) to our data (Wright and Charlesworth 2004), to test for the action of natural selection among candidate loci. We generated 1,000,000 cycles of the Markov chain (i.e., the chain length) assuming both neutral and selection models, to construct likelihood ratio tests to identify loci showing statistical support of selection —where twice the difference in log likelihood between the models is approximately chi-squared distributed. Again, divergence was estimated between species as well as to the inferred ANC sequence. The code and documentation are available for download at: http://www.yorku.ca/stephenw/Stephen_I._Wright/Programs.html.

As a separate approach aimed at identifying the fraction of positively selected amino acid fixations, we implemented the method of Eyre-Walker and Keightley (2009). Using information from both the SFS and divergence, this approach estimates both this proportion as well as a simple demographic model (by assuming that the population begins at equilibrium and experiences a step change in size t

generation ago). The fraction of advantageous amino acid divergence (α) is estimated as:

$$\alpha = \frac{d_N - d_S \int_0^\infty 2Nu(N,s)f(s|a,b)ds}{d_N},$$

where *f(s|a,b)*—the distribution of effects of deleterious mutations—is a gamma distribution with scale parameter *a* and shape parameter *b*. *N* is the effective population size, *u* the mutation rate per site, and thus, $2Nu(N,s)$ gives the rate of fixation from recurrent mutation. We use synonymous sites to define a neutral class (i.e., $s = 0$), and d_N and d_s are the numbers of selected (i.e., nonsynonymous) and neutral (i.e., synonymous) substitutions per site, respectively. The difference between the observed and expected (as determined from the neutral class) rate of selected substitution corresponds to the estimate of the proportion of adaptive substitutions. All necessary code for performing this analysis is available at: http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html.

Evaluating Models of Purifying Selection

To estimate the extent of purifying selection, we implement the approach of Loewe et al. (2006). This method was developed to characterize the fitness effects of deleterious nonsynonymous mutations, using polymorphism data from two species with different effective population sizes. Briefly, the underlying premise is that variants subject to sufficiently strong purifying selection will not increase significantly as effective population size increases, whereas neutral diversity is expected to increase proportionally with population size. Thus, the extent to which nonsynonymous diversity differs between species with different levels of synonymous site diversity should provide information regarding the strength of purifying selection. Thus, for species *i*, they define $\pi_{S_i} = 4N_{e_i}u$, $\pi_{A_i} = 4c_nN_{e_i}u + (1 - c_N)H_{P_i}$, $K_{S_i} = u$, and $K_{A_i} = c_nu + (1 - c_n)K_{P_i} + c_au$.

Here, H_{P_i} is the mean equilibrium diversity at sites subject to purifying selection, K_{P_i} is the mean substitution rate at these sites, c_N is the fraction of neutral nonsynonymous mutations, *u* is the expected mutation rate per site, and c_a measures the substitution as a fraction of all mutations.

Assuming a model of strong purifying selection ($N_es > 1$), the equilibrium diversity contributed by sites subject to purifying selection is well approximated by the deterministic expression $2u/s$ (McVean and Charlesworth 1999). Thus, one can simplify as $\pi_{A_i} = c_n\theta_i + 2(1 - c_n)\frac{u}{s_h}$, where $\theta_i = 4N_{e_i}u$, and s_h is the harmonic mean of selection coefficients (assumed to be the same in both species), and K_{P_i} becomes negligibly small. Thus, $K_{A_i} = c_nu + c_au$, and $c_n = \frac{\pi_{A_2} - \pi_{A_1}}{\pi_{S_2} - \pi_{S_1}}$. Substituting, we estimate selection as: $2N_{e_1}s_h = \frac{\pi_{S_1}(\pi_{A_1} + \pi_{S_2} - \pi_{A_2} - \pi_{S_1})}{\{\pi_{A_1}(\pi_{S_2} - \pi_{S_1}) - \pi_{S_1}(\pi_{A_2} - \pi_{A_1})\}}$, and $c_a = \frac{K_A}{K_S} - c_n$.

In order to account for the confounding effects of population history on the inference of purifying selection, Williamson et al. (2005) proposed a likelihood model–based approach in which data from a putatively neutral class (here synonymous sites) is estimated and fixed in order to perform the estimation of selection on the putatively selected class (nonsynonymous sites). As such, this approach also provides a demographic estimate (a stepwise size change at some time in the past), which may be compared with the above described approaches. Briefly, given that the expected number of polymorphic sites with *i* derived alleles segregating in a sample of *n* is $E[x_i] = \theta_1F_1(i, n; \tau, \upsilon)$, the probability that a particular single nucleotide polymorphism is at frequency *i* out of *n* is:

$$P_1(i,n;\tau,\upsilon) = \frac{F_1(i,n;\tau,\upsilon)}{\sum_{i=1}^{n-1} F_1(j,n;\tau,\upsilon)},$$

where υ = ancestral population size/current population size and τ = the time of the size change. With selection, we have the function:

$$F_2(i,n;\gamma,\tau,\upsilon) = \int_0^1 \binom{n}{i} q^i(1-q)^{n-i} f_2(q;\gamma,\tau,\upsilon)dq,$$

where there is the additional parameter $\gamma = 2Ns$, and the expected number of polymorphic sites segregating at a frequency *i* in a sample of size *n* becomes $E[x_i] = \theta_2F_2(i, n, \gamma, \tau, \upsilon)$. Thus, the probability that a particular polymorphic site is at frequency *i* out of *n* is:

$$P_2(i,n;\gamma,\tau,\upsilon) = \frac{F_2(i,n;\gamma,\tau,\upsilon)}{\sum_{i=1}^{n-1} F_2(j,n;\gamma,\tau,\upsilon)}.$$

Thus, to estimate the demographic parameters τ and υ, the likelihood function is maximized using class 1 data (synonymous sites). Then, for class 2 data (nonsynonymous sites), these parameters (τ and υ) are fixed in order to maximize the expression and estimate the selection parameter, γ. Thus, inherently, this approach does not account for the effects of linkage on synonymous sites.

Finally, the Eyre-Walker and Keightley (2009) approach described above also allows for estimation of parameters of deleterious mutations while additionally accounting for demography and the presence of beneficial mutations.

Results and Discussion

Patterns of Diversity and Estimating Species-Specific Demographic Models

In *D. miranda*, mean $\pi_{syn} = 0.006$, $\theta_{syn} = 0.007$, mean Tajima's (1989) $D_{syn} = -0.38$, and mean Fay and Wu's (2000) $H_{syn} = 0.10$, all suggesting a slight excess of rare variants. In *D. pseudoobscura*, mean $\pi_{syn} = 0.014$, $\theta_{syn} = 0.019$, mean Tajima's $D_{syn} = -0.37$, and mean Fay and Wu's $H_{syn} = 1.81$, suggesting both a similar excess of rares as well as a larger current effective population size (table 1). At nonsynonymous sites, for *D. miranda*, mean $\pi_{NS} = 0.0004$, $\theta_{NS} =$

Table 1
Summary Statistics of Synonymous (and nonsynonymous) Patterns of Variation in *Drosophila miranda* and *D. pseudoobscura*

	D. miranda	*D. pseudoobscura*
average *n*	14	14
π_{syn} (π_{NS})	0.006 (0.0003)	0.014 (0.0011)
θ_{syn} (θ_{NS})	0.007 (0.0004)	0.019 (0.0013)
Taj D_{syn} (D_{NS})	−0.38 (−1.1)	−0.37 (−0.26)
F&W H_{syn} (H_{NS})	0.10 (0.09)	1.81 (1.57)

0.0003, mean Tajima's (1989) $D_{NS} = -1.06$, and mean Fay and Wu's (2000) $H_{NS} = 0.089$. In *D. pseudoobscura*, mean $\pi_{NS} = 0.0014$, $\theta_{NS} = 0.0011$, mean Tajima's $D_{NS} = -0.26$, and mean Fay and Wu's $H_{NS} = 1.57$ (table 1).

Utilizing a recently proposed likelihood-based demographic estimator, dadi (Gutenkunst et al. 2009), we estimate demographic models for both *D. miranda* and *D. pseudoobscura*, using our large and consistently sampled data set. dadi infers demographic parameters by using a diffusion approach to fit the site-frequency spectrum of the observed data to a demographic model. Consistent with the conclusions of Bachtrog and Andolfatto (2006), we estimate a severe bottleneck for *D. miranda*. The estimated model begins with a much larger ancestral population size, followed by a reduction to 0.0005 of the ancestral size at 0.12 4*N* generations in the past, with the reduction lasting 0.02 4*N* generations. At this time, the population size recovers to 0.48 of the ancestral size. The relatively severe and long-lasting size reduction, followed by only moderate growth, results in a considerable reduction in diversity in *D. miranda*, relative to the ancestral population (fig. 1). In *D. pseudoobscura*, a very different demographic model is estimated. Although the best-fitting demographic model includes a relatively minor size reduction, the demographic history of *D. pseudoobscura* is mainly characterized by a large and relatively stable population size, which has recently experienced moderate growth (fig. 1). Specifically, the population size is estimated to have experienced a reduction to 0.81 of the ancestral size at 0.18 4*N* generations ago, lasting 0.09 4*N* generations. At this time, the population recovers to 1.35 of the ancestral size (i.e., growth). Taken together, these estimated models yield two important conclusions with regards to comparing effective population sizes between the two species: 1) in general, *D. pseudoobscura* and *D. miranda* may have had similar ancestral population sizes, and *D. pseudoobscura* has had a considerably more stable population history than *D. miranda* since the species split; and 2) *D. miranda* appears to have undergone a recent and severe size reduction, thus exaggerating the difference in their current effective population size.

Levels of synonymous polymorphism contain information about current effective population sizes (i.e., on average $4N_e$ generations ago) and suggest a roughly 3-fold difference in current N_e between species (bootstrap 95% CI = 2.3–4.1). Back-calculating from the dadi inference suggests an ancestral N_e of less than 2-fold difference (by assuming that the current population size of *D. pseudoobscura* is in fact 3-fold greater than *D. miranda*).

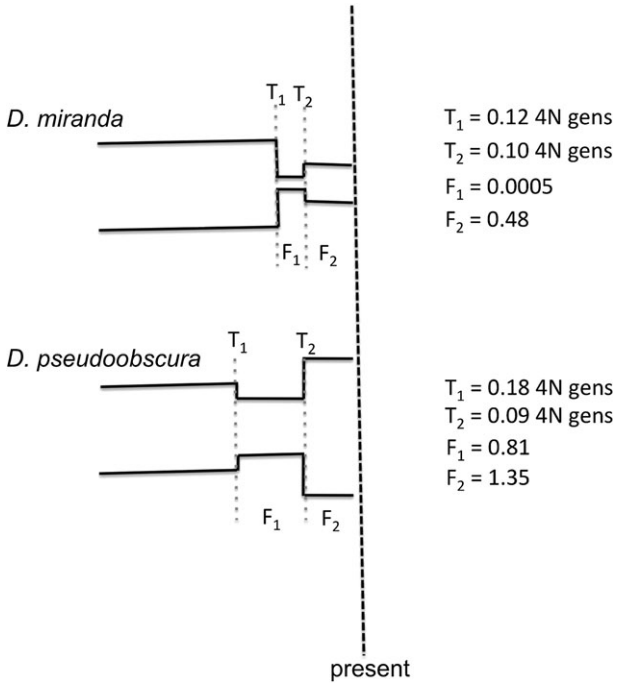

Fig. 1.—A cartoon schematic of the demographic models estimated for *D. miranda* and *D. pseudoobscura*. The demographic history of *D. miranda* is characterized by a severe bottleneck, while *D. pseudoobscura* is inferred to have had a relatively long-term stable population size, followed be recent growth. Thus, while the recent demographic history has served to exaggerate differences in size, the ancestral population size after the split of the two species may have been similar. T_1 and T_2 are the estimated times of the population size changes in 4*N* generations, and F_1 and F_2 are the changes in population size associated with the event (where *F* is the population size fraction relative to the ancestral size). The vertical black dotted line indicates the present time of sampling (*i.e.*, $t = 0$).

An alternative approach for estimating effective population size between species utilizes patterns of codon usage/bias (Bulmer 1991). By assuming that back and forward mutation rates are equal, it is possible to calculate selection coefficients from the proportion of optimal codon usage and transform this into an estimate of relative long-term effective population size (see calculations in Methods). The estimated ratio of population size of *D. miranda* versus *D. pseudoobscura* is 0.89 (bootstrap 95% CI = 0.79–0.93). Given that patterns of codon usage are expected to reach equilibrium very slowly, the more similar estimate in N_e for these two species appears consistent with the estimated demographic model using dadi, where the major size change between species has occurred only recently in evolutionary time. Thus, although current N_e may differ substantially between species, these calculations suggest that ancestral N_e may be much more similar (table 2). This implies that if

Table 2
Summary of Demographic Models Estimated Using Different Procedures

	Drosophila miranda	*D. pseudoobscura*	mir_{N_e}/pse_{N_e}
dadi[a](N_{anc} = 1)	Bottleneck: t_{bn} = 0.12 N_{bn} = 0.0005 Recovery: t = 0.10 $N_{current}$ = 0.48	Bottleneck: t_{bn} = 0.18 N_{bn} = 0.81 Recovery: t = 0.09 $N_{current}$ = 1.35	—
Codon usage[b]	—	—	0.89
DFE-α[c](N_{anc} = 1)	Growth: t_g = 0.37 $N_{current}$ = 6.0	Growth: t_g = 0.08 $N_{current}$ = 6.6	—
Williamson et al.[d](N_{anc} = 1)	Bottleneck: t_{bn} = 0.10 $N_{current}$ = 0.001	Bottleneck: t_{bn} = 0.12 $N_{current}$ = 0.71	—
Jensen et al.[e]	—		0.26

[a] Estimation procedure of Gutenkunst et al. (2009). Values indicate the reduction in variation at the time of the size change (e.g., population is reduced to 0.0005 of the ancestral size), the period of reduction in 4*N* generations (e.g., reduction at 0.12 4*N* generations, lasting 0.02 4*N* generations), and the size to which the population recovers after the reduction (e.g., population recovers to 0.48 of the ancestral size).
[b] Estimation procedure of Bulmer (1991), in which a relative size is estimated based on patterns of codon usage.
[c] Estimation procedure of Eyre-Walker and Keightley (2009) coestimated with the fraction of selected sites. A stepwise growth in population size is estimated (e.g., 6-fold growth relative to the ancestral size) at a given time (e.g., 0.37 4*N* generations in the past).
[d] Estimation procedure of Williamson et al. (2005), coestimated with the fraction of selected sites. A stepwise reduction in population size is estimated (e.g., reduction to 0.001 of the ancestral size) at a given time (e.g., 0.1 4*N* generations in the past).
[e] Estimation of neutral θ from the estimation procedure of Jensen, Thornton, and Andolfatto (2008)—the relative size is presented as estimated from patterns of polymorphism.

population size is influencing rates of adaptive evolution, we might expect that these two species have experienced similar rates of adaptation in the past whereas current rates of adaptation might be more different.

To estimate rates of current and historical selection, we apply a series of statistical tests to our data. However, because nonequilibrium demographic histories (such as those estimated above) can severely bias tests of selection (see for example, Thornton et al. 2007), a number of approaches are taken in the following sections to minimize this effect: 1) *P* values for tests of selection are explicitly corrected based on the inferred demographic histories; 2) some of the tests employed to estimate selection (i.e., the GOF and ω_{max} statistics) have been specifically proposed to be robust to demographic histories such as those estimated here; 3) methods are employed that allow for the estimation of the fraction of selected sites while coestimating a demographic model utilizing a class of neutral sites (Williamson et al. 2005; Eyre-Walker and Keightley 2009); 4) the method of Jensen, Thornton, and Andolfatto (2008) for estimating distributions of the strength and rate of recurrent hitchhiking was demonstrated to be largely robust to nonequilibrium perturbations; and 5) we perform forward simulations to explicitly model adaptive evolution using the demographic history inferred for both *D. miranda* and *D. pseudoobscura*, to directly address the question of how rates of adaptation (and statistical power to detect selection) are expected to differ for these two species.

Purifying Selection and Demography in *D. miranda* and *D. pseudoobscura*

Williamson et al. (2005) proposed an elaborate approach for quantifying the action of purifying selection that attempts to account for nonequilibrium demography. Specifically, by utilizing a putatively neutral class of sites (i.e., synonymous sites), a demographic model is first fit to the data and then selection on nonsynonymous sites is estimated under the inferred demographic model. By rescaling their estimated demographic parameters, it is possible to directly compare their inferred demographic model with the approaches described above. Although the Williamson et al. approach only estimates a simple population size change (ν) at some point in the past (τ), the results are roughly compatible with those obtained under the more complicated estimation procedure employed by dadi. *Drosophila miranda* is inferred to have experienced a reduction to 0.001 of the ancestral size at 0.10 4*N* generations ago, and *D. pseudoobscura* is inferred to have experienced a reduction to 0.71 of the ancestral size at 0.12 4*N* generations in the past (table 3). Fixing these parameters and maximizing the likelihood function (see Methods), we estimate purifying selection on amino acid mutations $2Ns = -1.32$ in *D. miranda* and -2.67 in *D. pseudoobscura*. This is consistent with the hypothesis of more efficient purifying selection acting on slightly deleterious amino acid variation in the larger species (table 4).

In addition, we implement the approach of Loewe et al. (2006) that also utilizes divergence data for estimating a model of purifying selection. The basic idea of this method is that whereas neutral diversity will increase proportionally with increasing population size, variation subject to strong purifying selection is expected to increase less rapidly with increasing population size. Thus, a comparison between these two classes of sites can provide information regarding the relative strength of purifying selection. Consistent with the Williamson et al. approach, we estimate roughly 2-fold stronger purifying selection acting on nonsynonymous mutations in *D. pseudoobscura*, with $2N_es = -1.95$ and $2N_es =$

Table 3
Estimated Demographic Models and the Fit to the Observed Synonymous Frequency Spectrum

Program	Command Line	Fit to Data[a]
dadi[b], mir	ms 20 100000 -t 10 -eN 0 0.48 -eN 0.1 0.0005 -eN 0.12 1	0.79
dadi, pse	ms 20 100000 -t 10 -eN 0 1.35 -eN 0.09 0.81 -eN 0.18 1	0.86
DFE[c], mir	ms 20 100000 -t 10 -eG 0.37 6	0.09
DFE, pse	ms 20 100000 -t 10 -eG 0.08 6.6	0.11
Williamson[d], mir	ms 20 100000 -t 10 -eN 0 0.001 -eN 0.1 1	0.54
Williamson, pse	ms 20 100000 -t 10 -eN 0 0.71 -eN 0.12 1	0.38

[a] Fraction of replicates within $\sigma = 0.01$ of empirically observed values of both mean Tajima's D_{syn} and π_{syn}.
[b] Gutenkunst et al. (2009).
[c] Eyre-Walker and Keightley (2006).
[d] Williamson et al. (2005).

−3.36 in *D. miranda* and *D. pseudoobscura*, respectively (table 4). The 95% confidence intervals are wide and in fact contain the Williamson et al. estimates (*D. miranda*: −0.9, −6.7; *D. pseudoobscura*: −2.4, −10.1).

We can also estimate parameters of purifying selection under nonequilibrium demography simultaneously with adaptive evolution (see below) using an approach recently described by Eyre-Walker and Keightley (2009). Because Eyre-Walker and Keightley, as well as Williamson et al., defines *s* as the selection coefficient against homozygotes for the deleterious allele, these values are one-half of that estimated by Loewe et al. because the methods assume semidominance. Under this approach, the DFE of amino acid mutations is estimated by maximum likelihood based on their site-frequency spectrum and that of sites assumed to be evolving neutrally (synonymous sites). Demographic changes are modeled by a single step change in size from N_1 to N_2 t generations in the past. Applying this approach to our data, we estimate that both species underwent recent population growth, with *D. miranda* having grown 6-fold at 0.37 $4N$ generations in the past and *D. pseudoobscura* having grown 6.6-fold at 0.08 $4N$ generations ago (table 2). The demographic model estimated using the Eyre-Walker and Keightley method differs considerably from the other models estimated (table 2). On one hand, this approach only models a single stepwise change in population size and thus may not be fitting the data as precisely as the multiparameter dadi approach. On the other hand, the Eyre-Walker and Keightley approach is simultaneously fitting a demographic and selection model to the data, whereas dadi estimates the demographic history ignoring natural selection. Notably, recent studies have questioned the accuracy of frequency spectrum–based approaches such as those used by Gutenkunest et al. and Williamson et al. (Myers et al. 2008).

Table 4
Estimated Selection Parameters across Both Species for the Multilocus Polymorphism-and Divergence-Based Recurrent Selection Statistics Used Here

	Drosophila miranda	*D. pseudoobscura*
s[a]	2×10^{-3}	9×10^{-4}
$2N\lambda$[b]	1×10^{-4}	5×10^{-3}
α[c]	0.78	0.83
$2Ns$[d,e]	−1.32(−1.95)	−2.67(−3.36)

[a] Mean selection coefficient; estimation procedure of Jensen, Thornton, and Andolfatto (2008).
[b] Mean rate of adaptation; estimation procedure of Jensen, Thornton, and Andolfatto (2008).
[c] Fraction of positively selected loci; estimation procedure of Eyre-Walker and Keightley (2009).
[d] Strength of purifying selection acting on nonsynonymous sites; estimation procedure of Williamson et al. (2005).
[e] Strength of purifying selection acting on nonsynonymous sites; estimation procedure of Loewe et al. (2006).

To estimate the fit of the demographic model obtained from different methodologies, we performed coalescence simulations using the program ms (Hudson 2002). Specifically, we simulated 100,000 neutral genealogies with the demographic parameters identified under each model and estimated which fraction of simulations are compatible with the observed values for both mean Tajima's D_{syn} and π_{syn}. Our simulation results suggest superior data fitting of the demographic model identified by dadi relative to the approach of Eyre-Walker and Keightley, which performs rather poorly (table 3). Note, however, that the simulations are performed ignoring positive and negative selection, and the accuracy of existing methods to infer demographic parameters relative to one another in the presence of both positive and negative selection, and the impact of differing assumptions made in different approaches, remains a topic in need of more thorough investigation.

Using the Eyre-Walker and Keightley method, we can infer the DFE of newly arising amino acid mutations for both *D. miranda* and *D. pseudoobscura* (fig. 2). Consistent with a smaller N_e in *D. miranda*, a larger fraction of newly arising synonymous mutations are under weaker purifying selection in this species (i.e., $1 < N_e s < 100$; see fig. 2). Thus, although the inferred demographic model and parameters of purifying selection differ somewhat between approaches, we generally find that the strength of purifying selection is reduced in *D. miranda*, as expected based on its smaller effective population size.

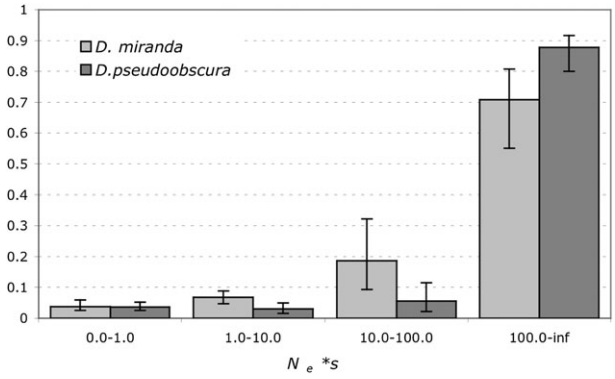

Fig. 2.—Distribution of fitness effects of newly arising amino-acid mutations estimated in the *D. miranda* and *D. pseudoobscura* lineages. Consistent with a smaller N_e in *D. miranda*, a larger fraction of newly arising nonsynonymous mutations are estimated to be under weak purifying selection (*i.e.*, $1 < N_e s < 100$).

Polymorphism-Based Inference of Positive Selection in *D. miranda* and *D. pseudoobscura*

We used several test statistics that identify adaptive evolution at individual loci utilizing different aspects of patterns of polymorphism (the CLRT, GOF, and ω_{max} tests). As stated above, these test statistics should have power to detect relatively recent adaptive evolution in the genome. In general, there is little evidence for positive selection in the frequency spectrum of *D. miranda*, with only marginal levels of rejection across statistics (table 5). In *D. pseudoobscura*, however, there are roughly 5- to 10-fold more loci that show significant evidence of selection than in *D. miranda* across statistics after a multiple test correction (table 5). This result is consistent with the expectation of a greater rate of adaptation in *D. pseudoobscura* due to its larger population size. However, given the severe size reduction estimated in *D. miranda*, frequency spectrum patterns associated with recent adaptation may have been eliminated by diversity-reducing bottleneck effects, resulting in less power to identify individual loci undergoing adaptive hitchhiking events (see power simulations below).

In addition, we also employ a multilocus method to infer parameters of adaptation in the two species. The method of Jensen, Thornton, and Andolfatto (2008) can estimate distributions of s and $2N\lambda$ under a recurrent hitchhiking model and has been shown to result in accurate estimation for data

Table 5
The Number of Significant Test Rejections across Both Species, after a Multiple Test Correction, for the Single-Locus Polymorphism and Divergence-Based Statistics Used Here

	Drosophila miranda	*D. pseudoobscura*
CLRT	3	27
GOF	1	11
ω_{max}	1	4
MK	2	18
HKA	6	12

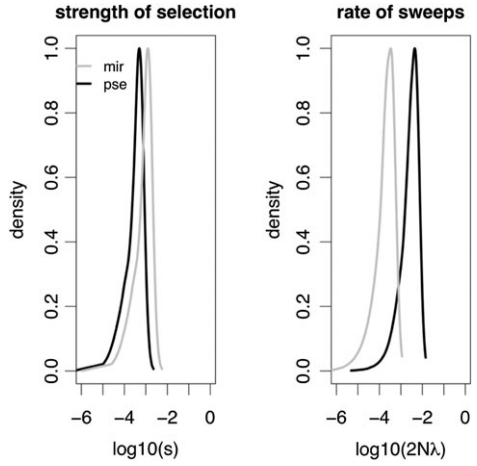

Fig. 3.—Approximate Bayesian estimation of both the strength and rate of recurrent positive selection, for randomly selected homologous genes from *D. miranda* and *D. pseudoobscura*. Estimation is based on 10^6 draws from the prior. Given are the marginal distributions, with *D. pseudoobscura* in black and *D. miranda* in gray. Consistent with an important role of effective population size driving adaptive evolution, roughly an order of magnitude greater rate of fixation is estimated for the currently larger *D. pseudoobscura* population using this polymorphism-based statistic.

sets of this size. Applying this method to our data, we infer maximum a posteriori (MAP) estimates of the mean $s = 2 \times 10^{-3}$ and mean $2N\lambda = 1 \times 10^{-4}$ for *D. miranda* and a mean $s = 9 \times 10^{-4}$ and mean $2N\lambda = 5 \times 10^{-3}$ for *D. pseudoobscura* (fig. 3). Thus, although the distribution of the strength of selection is similar between species, there is a significant shift in the distribution of the rate of selection. Specifically, the rate of recurrent hitchhiking is estimated to be roughly an order of magnitude greater in *D. pseudoobscura* compared with *D. miranda*. Thus, consistent with the expectation of differences in current N_e between species, this polymorphism-based approach that captures recent selective events indicates a considerable difference in rates of adaptive evolution between species (table 4). Consistent with other polymorphism-based estimators, a large difference in population size is estimated between the two species, with an estimated $N_e = 1.15^6$ for *D. miranda* and $N_e = 4.50^6$ for *D. pseudoobscura*, or a relative ratio of 0.26 (table 2).

Importantly, as we are estimating the joint parameter of effective population size and the rate of selection, we can determine whether the estimated difference in population size alone is sufficient to account for the estimated rate difference. If we use the coalescent effective population size estimated from dadi, $\lambda = 5.0 \times 10^{-10}$ and 1.3×10^{-9} in *D. miranda* and *D. pseudoobscura*, respectively; whereas for the estimated current population size from polymorphism data, $\lambda = 4.1 \times 10^{-11}$ and 9.7×10^{-10}, respectively. Thus, although population size explains a good deal of the estimated difference in $2N\lambda$, there does appear to be

Table 6
Summary of Forward Simulation Results

	Drosophila miranda	*D. pseudoobscura*
$Severity_{bneck}$[a]	0.0005	0.81
$Time_{bneck}$[b]	0.12	0.18
$Duration_{bneck}$[c]	0.02	0.09
$Recovery_{bneck}$[d]	0.48	1.35
s[e]	9×10^{-4}	9×10^{-4}
$2N\lambda$[f]	1×10^{-4}	5×10^{-3}
MAP $2N\lambda$[g]	**8×10^{-4}**	**1×10^{-3}**
Power (CLRT/GOF)[h]	0.38	0.81

[a] Severity of the simulated reduction in population size relative to the ancestral size.
[b] Time of the simulated reduction in population size in $4N$ generations.
[c] Duration of the simulated reduction in population size in $4N$ generations.
[d] Simulated size to which the population recover postbottleneck, relative to ancestral size.
[e] Simulated selection coefficeint.
[f] Simulated rate of fixation.
[g] Estimated rate of fixation for the above parameters.
[h] Power of the CLRT/GOF combination for the above parameters.

a consistently larger λ in *D. pseudoobscura*. This indicates a greater rate of fixation of beneficial mutations in the larger *D. pseudoobscura* population, possibly because a larger fraction of slightly beneficial mutations—although effectively neutral in *D. miranda*—are subject to positive selection. This is consistent with the slightly smaller estimate of *s* in *D. pseudoobscura* (fig. 3).

In order to evaluate concerns regarding both reduced power to detect selection in the severely bottlenecked *D. miranda* population as well as to evaluate hypotheses regarding the expected differences in rates of adaptation in these two nonequilibrium populations, we performed a series of forward simulations (see Methods). In particular, we simulated demographic models as estimated by dadi with selection parameters estimated from the recurrent hitchhiking model ($s = 9 \times 10^{-4}$ and mean $\lambda = 5.0 \times 10^{-10}$). Using these parameter assumptions, the simulated population adaptive fixation rate in *D. pseudoobscura* is much faster than in *D. miranda* ($2N\lambda = 1 \times 10^{-4}$ for *D. miranda* and 5×10^{-3} for *D. pseudoobscura*). We applied the approximate Bayesian framework of Jensen, Thornton, and Andolfatto (2008) to estimate parameters of recurrent hitchhiking (RHH) from these simulated data sets. Consistent with the results of Jensen, Thornton, and Andolfatto (2008), this method is generally robust to demography (MAP estimates of $2N\lambda = 8 \times 10^{-4}$ and 1×10^{-3} [table 6], in *D. miranda* and *D. pseudoobscura*, respectively). The increased variance created by the severe bottleneck in *D. miranda* is elevating the estimated rate of adaptive evolution somewhat, whereas the variance reducing effect of population growth in *D. pseudoobscura* is resulting in a slight underestimate. Thus, the simulation results suggest that if a bias is being created by the underlying demographic histories of these species, the likely result is a homogenizing effect on estimated rates of adaptation between species (i.e., contrary to empirical observation). These simulations also serve as an effective bootstrap for determining statistical significance between the estimated distributions. For the strength of selection, the MAP estimate for *D. miranda* is contained within the 95% credible interval (CI) of *D. pseudoobscura*, and vice versa, whereas for the rate of selection, the MAP estimates are not contained within the CIs of the opposing species.

In order to further evaluate the hitchhiking model, we use forward simulations. Given that the expected waiting time (in $4N$ generations) between beneficial fixations is $1/M \times 2N\lambda$, where M is the size of the sampled region—the waiting time is relatively shorter in *D. pseudoobscura* owing to its larger N (for $M = 100$ kb, the expected time between fixations = 0.05 in *D. miranda* and 0.0025 in *D. pseudoobscura*). Consistent with previous results, the power to detect selection using single-hitchhiking model–based statistics (where the assumption is that the beneficial mutation has reach fixation immediately prior to sampling), under recurrent hitchhiking is poor (Przeworski 2002; Jensen, Thornton, and Aquadro 2008; Jensen 2009). For the CLRT/GOF combination, the power to detect hitchhiking events in *D. miranda* was found to be 0.38, and 0.81 in *D. pseudoobscura* (table 6)—where power is assessed as the fraction of replicates rejecting the neutral model in the CLRT and being consistent with a hitchhiking model in the GOF test. As opposed to RHH estimation where a bias induced by demography may be bringing estimates of the rate of adaptation nearer to one another, single hitchhiking (SHH) approaches to detect selection have reduced power in the species with the smaller effective population size. Thus, although some of the difference in rejections of SHH models can be explained by differences in power (i.e., roughly 2-fold between species), this may not be sufficient to account for the empirical observation of 10-fold more rejections between *D. pseudoobscura* and *D. miranda* (CLRT = 27 vs. 3 rejections, GOF = 11 vs. 1). Thus, consistent with the inferred difference in current N_e between species, we consistently infer higher rates of adaptive evolution in the larger *D. pseudoobscura* population.

Divergence-Based Inference of Selection in *D. miranda* and *D. pseudoobscura*

Divergence-based approaches to estimate rates of adaptation yield information about the action of selection over a longer time period (i.e., since the split of the two species). Thus, much of the adaptation detected using divergence data might have in fact occurred in an ANC whose population history differs substantially from that of the current population. In contrast, polymorphism-based approaches can only detect selection on a much more recent timescale (i.e., within the population coalescent time, most estimators only have reasonable power to detect selection as recent as 0.1 $4N_e$ generations ago; Przeworski 2002). Thus, given the complicated demographic history of both *D. miranda* and *D. pseudoobscura*, we might expect polymorphism- and divergence-based approaches of selection to yield different

conclusion about relative rates of adaptation in the two species. Specifically, given the longer timescale over which divergence-based estimators can detect selection, together with the much more similar ancestral population sizes estimated for the two species—we may expect that divergence-based estimates of adaptation are more similar between species. Conversely, given the much larger difference in estimated current population sizes, combined with the severe bottleneck estimated in *D. miranda*, polymorphism-based estimates of selection may differ more dramatically between species.

A variety of population genetics approaches exist to quantify adaptive evolution utilizing sequence divergence between species. Two of the most widely used approaches for simultaneously considering polymorphism and divergence data are the MK and HKA tests (see Methods). Applying these test statistics to our data, we generally find a greater proportion of positively selected loci in *D. pseudoobscura*. However, the disparity in rates of adaptive evolution between species is not as great as with polymorphism-based statistics. Consistent with our expectations based upon the estimated demographic model, only a 2- to 4-fold greater proportion of loci show significant evidence of selection in the larger *D. pseudoobscura* population (as opposed to 5- to 10-fold; table 5). This significant result holds both in the presence and absence of a reconstructed ANC. The recent and severe size reduction estimated for *D. miranda*, combined with more similar divergence-based estimates of adaptive evolution, appears consistent with a larger ancestral population size for *D. miranda* and thus a more similar rate of adaptation to *D. pseudoobscura* over a significant portion of the species history. Conversely, the recent bottleneck in *D. miranda* increases their difference in effective population size, thus creating a greater disparity in polymorphism-based statistics to detect adaptation.

To estimate the fraction of amino acid mutations driven to fixation by positive selection and simultaneously coestimate a demographic model, we implemented the approach of Eyre-Walker and Keightley (2009). In *D. miranda*, the estimated fraction of advantageous amino acid mutations is 0.78, and 0.83 in *D. pseudoobscura* (table 4). Again, this divergence-based estimate suggests similar rates of adaptive amino acid evolution for this species pair. Additionally, we also calculate lineage-specific estimates of α, the fraction of adaptive amino acid evolution, in *D. miranda* and *D. pseudoobscura*, using different approaches based on the MK-framework (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004). If all polymorphic sites are used for estimation, we find α to be consistently lower in the *D. miranda* lineage (though not significantly, table 7). However, it is well known that segregating deleterious amino acid mutations lead to biased estimates of α in this type of analysis, and more slightly deleterious amino acid mutations appear to be segregating in *D. miranda* (see fig. 2). A commonly used procedure to remedy this effect is to exclude low-frequency mutations. Indeed, if we only consider polymorphisms at a frequency above 10%, estimates of α are almost identical between the two species, indicating very similar lineage-specific rates of adaptive amino acid evolution.

Table 7
ineage-Specific Estimates of α, the Fraction of Adaptive Amino Acid Substitutions (and 95% confidence intervals), Using *Drosophila affinis* or *D. athabasca* as a Second Outgroup

Method	*D. pseudoobscura*	*D. miranda*
All sites		
α^a	0.66 (0.55–0.74)	0.57 (0.38–0.70)
α^b	0.57 (0.43–0.67)	0.45 (0.23–0.61)
α^c	0.56 (0.43–0.66)	0.49 (0.29–0.64)
$f > 0.1^d$		
α^a	0.72 (0.63–0.80)	0.70 (0.55–0.80)
α^b	0.68 (0.58–0.76)	0.69 (0.54–0.79)
α^c	0.65 (0.52–0.75)	0.69 (0.53–0.80)

[a] Fraction of adaptive amino acid mutations; estimation procedure of Fay et al. (2001).
[b] Fraction of adaptive amino acid mutations; estimation procedure of Smith and Eyre-Walker (2002).
[c] Fraction of adaptive amino acid mutations; estimation procedure of Bierne and Eyre-Walker (2004).
[d] Fraction of adaptive amino acid mutations ignoring polymorphism at a frequency < 0.1.

Although the demographic model estimated under the Eyre-Walker and Keightley scheme is simplified compared with the dadi procedure (inasmuch as it is restricted to a step change in population size from N_1 to N_2, t generations ago similar to the procedure of Williamson et al. 2005), demography is effectively coestimated with a selection model. For both species, a growth model is coestimated (see above), whereas dadi infers a bottleneck in both species, followed by growth. If selection is indeed widespread across the genome, as suggested by our results, this discrepancy between methods may be expected. In particular, dadi is estimating a purely neutral model, and it may be forced to account for the diversity-reducing and frequency spectrum-skewing effects of recurrent hitchhiking under neutrality. Preliminary analysis from forward simulation indeed suggests that dadi is biased in the direction of estimating bottlenecks of increased severity and duration, as the fraction of positively selected loci increases. Thus, although analyses consistently point to a more similar ancestral size between the two species and a large difference in current population sizes—incorporating selection into the demographic estimation procedure suggests that the nonequilibrium history may not be as severe as the neutral demographic model may suggest.

Current and Historical Selection in *D. miranda* and *D. pseudoobscura*

Consistent with previous observations in Drosophila (e.g., Andolfatto 2007), a significantly negative correlation is observed between K_a and π_s in both species (i.e., levels of synonymous site diversity are reduced in genes with rapid amino acid evolution; fig. 4). Interpreting this pattern in

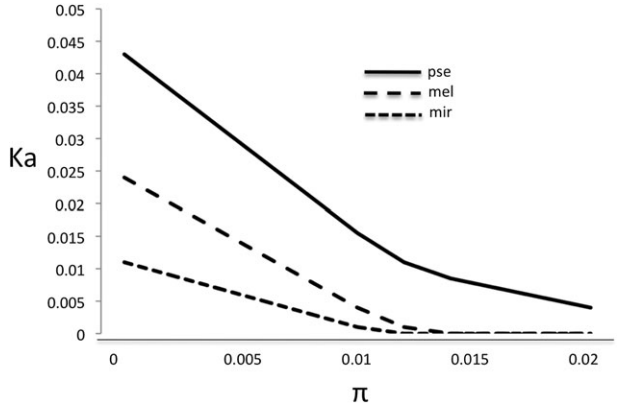

Fig. 4.—Plot of p_s versus K_a for *D. miranda* (the pooled loci of both this study, as well as the randomly selected genes of Bachtrog *et al.* 2009), *D. pseudoobscura* (the pooled loci of homologous genes across the X-chromosome) and *D. melanogaster* (Bachtrog 2008). The solid line indicates the significant correlation between these measures of synonymous polymorphism and non-synonymous divergence — a prediction consistent with both genetic hitchhiking and background selection - among these species of differing effective population sizes (with *D. melanogaster* thought to be of intermediate size between *D. miranda* and *D. pseudoobscura*).

isolation has proven difficult because models of both positive and negative selection can, in principle, produce this correlation. Specifically, recurrent fixations of advantageous amino acid mutations can each contribute to local reduction in neutral variation due to hitchhiking effects (Maynard Smith and Haigh 1974; Kaplan et al. 1989). Conversely, models of background selection (i.e., the removal of weakly deleterious mutations) can result in local reductions of N_e (Charlesworth et al. 1993). This, in turn, results in reduced levels of neutral diversity and also decreases the efficiency of purifying selection, thereby potentially causing an accumulation of weakly deleterious amino acid mutations (i.e., reducing π_s and elevating K_a; Charlesworth 1994).

Interestingly, for both *D. miranda* and *D. pseudoobscura*, we observe a significantly negative association between CLRT *P* values and K_a (fig. 5). Thus, genes that show higher rates of amino acid divergence show more statistical evidence of recent positive selection at the polymorphism level. Given that the CLRT test is robust to frequency spectrum perturbations caused by background selection (Kim and Stephan 2002), the correlation between K_a and the CLRT *P* values suggests that neutral polymorphism at rapidly evolving genes in Drosophila is, at least partially, influenced by recurrent positive selection. This correlation further suggests that, even between these species with different current population sizes, selection is frequent enough to create a significant relationship between polymorphism- and divergence-based comparisons of selection in both species. This also indicates that many genes that have been evolving adaptively in the more distant past (and thus have elevated K_a) are still undergoing adaptive evolution in both *D. miranda* and *D. pseudoobscura*.

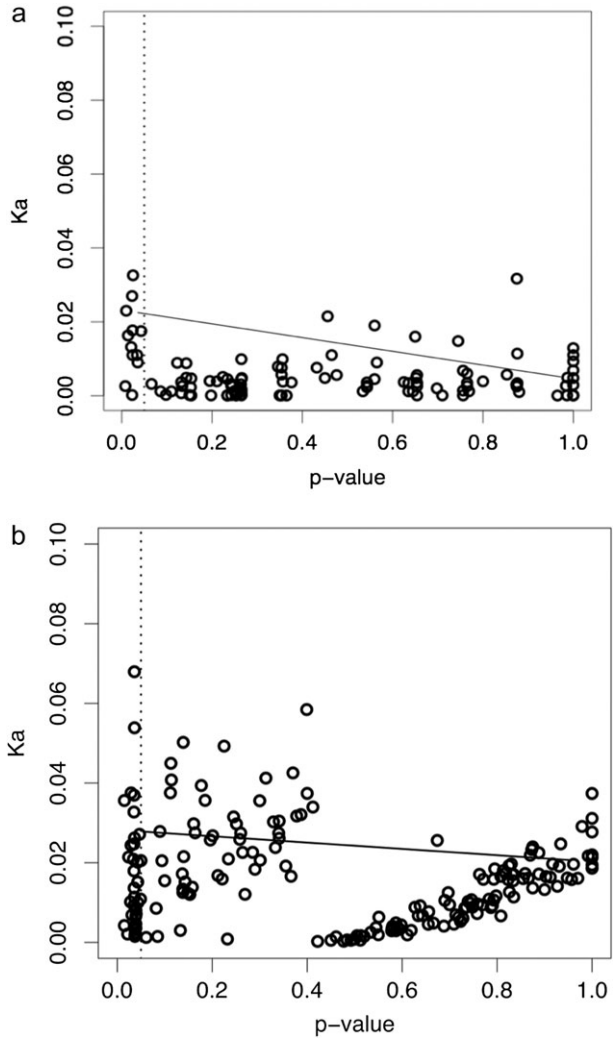

Fig. 5—Plot of CLRT *P* values with K_a. (*A*) *Drosophila miranda*: The pooled loci of both this study, as well as the randomly selected genes of Bachtrog et al. (2009), are shown. (*B*) *Drosophila pseudoobscura*: The pooled loci of homologous genes across the X chromosome. The dotted line indicates the 5% significance cutoff for the CLRT. The solid line indicates the significant correlation between the observed divergence measure K_a and the calculated *P* value of this polymorphism-based test statistic. Results indicate a significant correlation between this polymorphism-based test of selection and this divergence-based measure, in both species—despite roughly an order of magnitude difference in effective population size. This result suggests that at least a portion of the correlation observed in figure 4 owes to hitchhiking effects.

Conclusions

Here, we present a randomly subsampled screen of over 100 genes in *D. pseudoobscura*, mirroring a data set previously produced for *D. miranda* (Bachtrog et al. 2009). Utilizing this large and consistently sampled data set, demographic models for both species are estimated. *Drosophila miranda* is characterized by a recent severe size reduction, whereas

the *D. pseudoobscura* population appears relatively stable since the species split, with recent moderate growth. This implies that although current population sizes differ substantially between these two species, their ancestral sizes are more similar. Consistent with an important role of effective population size driving patterns of adaptation, we consistently infer higher rates of positive selection in the larger *D. pseudoobscura* population. Although more beneficial mutations occur each generation in a larger population, this also suggests that a substantial fraction of newly arising beneficial mutations are effectively neutral in the smaller *D. miranda* population, but under selection in *D. pseudoobscura*.

Our study also highlights important differences between polymorphism- and divergence-based estimators of recurrent selection models, and their interaction with the species' underlying demographic history. Consistent with the estimated model of a historically more similar ancestral population size, divergence-based approaches to detect selection suggest rather similar rates of adaptation for both species. Conversely, polymorphism-based approaches suggest a much more prevalent role for selection shaping patterns of genomic variation in *D. pseudoobscura*, consistent with the inferred recent size reduction in *D. miranda* and recent growth in *D. pseudoobscura*. This discrepancy can be understood in relation to the relative timescales for which these different classes of test statistics are sensitive to detect selection.

Finally, consistent with the recent results of Haddrill et al. (2010), evidence suggests pervasive roles for both positive and purifying selection—creating significant correlations between polymorphism- and divergence-based methodologies, and being generally consistent with the Darwin domain of molecular evolution. Our study demonstrates that the comparison between both polymorphism- and divergence-based approaches, coupled with demographic estimates, may provide a much more comprehensive view of adaptation.

Acknowledgments

We thank Nick Toda for contributing to data generation. J.D.J. is supported by National Science Foundation grant DEB-1002785 and a Worcester Foundation award. D.B. is supported by National Institutes of Health grant GM076007, a Sloan Research Fellowship, and a David and Lucille Packard Fellowship.

Literature Cited

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 17:1755–1762.

Bachtrog D. 2008. Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. BMC Evol Biol. 8:334.

Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in Drosophila miranda. Genetics 174:2045–2059.

Bachtrog D, Gordo I. 2004. Adaptive evolution of asexual populations under Muller's ratchet. Evolution 58:1403–1413.

Bachtrog D, Jensen JD, Zhang Z. 2009. Accelerated adaptive evolution on a newly formed X chromosome. PLoS Biol. 7:e82.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol. 27:1350–1360.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Bustamante CD, et al. 2002. The cost of inbreeding in Arabidopsis. Nature 16:531–534.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res. 63:213–227.

Charlesworth B, Bartolome C, Noel V. 2005. The detection of shared and ancestral polymorphisms. Genet Res. 86:149–157.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Charlesworth D, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 23:1348–1356.

Clark AG. 1997. Neutral behavior of shared polymorphism. Proc Natl Acad Sci U S A. 94:7730–7734.

Crow JF, Kimura M. 1970. An introduction to population genetic theory. Edina, MN: Alpha editions.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol. 21:569–575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26:2097–2108.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics 158:1227–1234.

Gillespie JH. 1999. The role of population size in molecular evolution. Theor Popul Biol. 55:145–156.

Gillespie JH. 2000. Genetic drift in an infinite population. The pseudohitchiking model. Genetics 155:909–919.

Gillespie JH. 2001. Is the population size of a species relevant to its evolution? Evolution 55:2161–2169.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

Haddrill PR, Loewe L, Charlesworth B. 2010. Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. Genetics 185:1381–1396.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2009. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6(1):e1000825.

Hernandez R. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24:2786–2787.

Hudson R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Jensen JD. 2009. On reconciling single and recurrent hitchhiking models. Genome Biol Evol. 1:320–324.
Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–1410.
Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests strong recurrent selective sweeps in Drosophila. PLoS Genet. 4(9):e1000198.
Jensen JD, Thornton KR, Aquadro CF. 2008. Inferring selection in partially sequenced regions. Mol Biol Evol. 25:438–446.
Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. Genetics 176:2371–2379.
Kaplan NL, Hudson RR, Langley CH. 1989. The 'hitchhiking effect' revisited. Genetics 123:887–899.
Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524.
Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777.
Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624.
Kimura M. 1983. Rare variant alleles in the light of the neutral theory. Mol Biol Evol. 1:84–93.
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2:e166.
Loewe L, Charlesworth B, Bartolome B, Noel V. 2006. Estimating selection on nonsynonymous mutations. Genetics 172:1079–1092.
Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics 177:2083–2099.
Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–25.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.
McVean GA, Charlesworth B. 1999. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155:929–944.
Myers S, Fefferman C, Patterson N. 2008. Can one learn history from the allelic spectrum. Theor Popul Biol. 73:342–348.
Nielsen R, et al. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.
Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol. 20:1231–1239.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189.
Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS Genet. 7(2):e1001302.
Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. J Mol Evol. 57:S154–S164.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome? PLoS Genet. 5:e1000495.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024.
Stephan W, Song Y, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172:2647–2663.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.
Thornton KR. 2009. Automating approximate Bayesian computation by local linear regression. BMC Genet. 10:35.
Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity 98:340–348.
Wiehe TH, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol Biol Evol. 10:842–854.
Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 102:7882–7887.
Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.
Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yi S, Bachtrog D, Charlesworth B. 2003. A survey of chromosomal and nucleotide sequence variation in *Drosophila miranda*. Genetics 164:1369–1381.

Associate editor: Brandon Gaut