Copyright © 2002 by the Genetics Society of America

Testing Models of Selection and Demography in *Drosophila simulans*

Jeffrey D. Wall,*[,1] Peter Andolfatto† and Molly Przeworski‡[,2]

**Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, †Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and ‡Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom*

Manuscript received December 10, 2001
Accepted for publication June 12, 2002

ABSTRACT

We analyze patterns of nucleotide variability at 15 X-linked loci and 14 autosomal loci from a North American population of *Drosophila simulans*. We show that there is significantly more linkage disequilibrium on the X chromosome than on chromosome arm 3R and much more linkage disequilibrium on both chromosomes than expected from estimates of recombination rates, mutation rates, and levels of diversity. To explore what types of evolutionary models might explain this observation, we examine a model of recurrent, nonoverlapping selective sweeps and a model of a recent drastic bottleneck (*e.g.*, founder event) in the demographic history of North American populations of *D. simulans*. The simple sweep model is not consistent with the observed patterns of linkage disequilibrium nor with the observed frequencies of segregating mutations. Under a restricted range of parameter values, a simple bottleneck model is consistent with multiple facets of the data. While our results do not exclude some influence of selection on X *vs.* autosome variability levels, they suggest that demography alone may account for patterns of linkage disequilibrium and the frequency spectrum of segregating mutations in this population of *D. simulans*.

A fundamental question in population genetics is the relative importance of natural selection *vs.* neutral and/or demographic factors in shaping genome-wide patterns of sequence variability (Lewontin 1974; Kimura 1983). Distinguishing between selective and neutral/demographic effects on genome variability requires multiple independent loci scattered throughout the genome. Although there are polymorphism data from dozens of loci in Drosophila (see, *e.g.*, Moriyama and Powell 1996; Andolfatto and Przeworski 2000; Andolfatto 2001; Przeworski *et al.* 2001), these data are difficult to interpret because many of the loci are sampled in different populations. As a result, one is never quite sure to what extent patterns of variation at a collection of loci are the byproduct of the particular sampling schemes used. In particular, in the absence of independent knowledge about the demographic history of a species, it becomes very difficult, if not impossible, to draw inferences about the role of natural selection. Much of this ambiguity can be eliminated by sequencing the same lines from the same populations at multiple loci, yet it is only recently that consistently sampled Drosophila data have been gathered from many loci (Begun and Whitley 2000; Begun *et al.* 2000; Andolfatto and Przeworski 2001).

In one such study, Begun and Whitley (2000) collected polymorphism data from 15 X-linked loci and 14 loci on chromosome arm 3R in a California population of *Drosophila simulans*. Their goal was to distinguish between different explanations for the empirical observation that levels of variability are positively correlated with rates of meiotic crossing over in Drosophila (Aguadé *et al.* 1989; Begun and Aquadro 1992; Aquadro *et al.* 1994). Two major theories have been proposed for this pattern; both describe the effects of natural selection at linked sites. The hitchhiking or selective sweep model posits that alleles driven to fixation by positive selection reduce levels of variation at linked neutral sites (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989), while the background selection model considers the variation-reducing effect of strong purifying selection (Hudson and Kaplan 1995; Charlesworth *et al.* 1993). Both models predict a greater reduction of variability in areas of reduced recombination, so both are potential explanations for the positive correlation between diversity levels and recombination rates.

Begun and Whitley (2000) attempt to distinguish between background selection and positive selection models by comparing levels of variability on the X chromosome and an autosome (Aquadro *et al.* 1994). If deleterious mutations tend to be partially recessive (Crow and Simmons 1983; Houle *et al.* 1997), purifying selection is expected to be more efficient on the X chromosome relative to the autosomes, due to haploidy in males. The background selection model therefore predicts that, if all other factors are comparable, there should be less of a reduction in levels of variation on the X chromosome than on the autosomes (Charles-

[1] *Corresponding author:* Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138. E-mail: jwall@fas.harvard.edu

[2] *Present address:* Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany.

worth *et al.* 1993; Charlesworth 1996). Begun and Whitley (2000), however, find reduced levels of diversity on the X chromosome, even when X-linked diversity is multiplied by $\frac{4}{3}$ (to correct for the fewer number of X chromosomes, assuming equal male and female effective population sizes). They conclude that the background selection model is incompatible with their data and suggest instead that greater hitchhiking effects due to positive selection on the X relative to the autosomes might account for their pattern. This will occur if positively selected alleles are sometimes recessive; haploidy on the X in males would then facilitate X-linked selective sweeps.

Considering what we know about the demographic history of *D. simulans* populations, it may also be relevant to consider nonselective explanations for Begun and Whitley's observation. *D. simulans* is a human commensal; it is thought to have originated in tropical Africa and may have colonized the Americas as recently as a few hundred years ago (David and Capy 1988; Lachaise *et al.* 1988). A contraction in population size (*i.e.*, a population bottleneck or founder event) during the initial colonization may have had a large impact on the patterns of variation in samples from migrant populations. Indeed, variability in New World populations appears to be lower than in African populations, as expected after a population size contraction (Hamblin and Veuille 1999; Andolfatto 2001). In addition, the average Tajima's *D* (a commonly used summary of the frequency spectrum of segregating mutations, *cf.* Tajima 1989a) at two X-linked loci is higher in non-African than in African populations; this also is suggestive of a population contraction, with an associated loss of low frequency alleles in migrant populations (Fay and Wu 1999; Hamblin and Veuille 1999).

In this article, we revisit Begun and Whitley's data, by explicitly considering both a model of recurrent selective sweeps and a recent bottleneck model; we also examine additional aspects of the data besides levels of diversity. Braverman *et al.* (1995) showed that a model of recurrent selective sweeps predicts a sharp excess of rare variants relative to the expectations of the standard constant-sized Wright-Fisher neutral model (called hereafter the null model). In our analyses, we study both *D* (Tajima 1989a) and a summary of linkage disequilibrium (*i.e.*, the nonrandom association between alleles at different nucleotide sites). We measure linkage disequilibrium by estimating (under the null model) the population recombination parameter ρ ($= 4Nr$, where N is the effective population size and r is the sex-averaged recombination rate per base pair per generation) from the sequence data at each locus (C_{HRM}, *cf.* Wall 2000). The parameter r is estimated from a comparison of *D. simulans* physical and genetic maps; we can then estimate N as $\hat{N}_\rho = \hat{\rho}/4\hat{r}$. Low estimated values of $\hat{N}_\rho$ indicate high levels of linkage disequilibrium and vice versa. A similar approach was used by Andolfatto and Przeworski (2000) using data from *D. melanogaster* and *D. simulans*. They found that N estimated from linkage disequilibrium and $\hat{r}$ was much smaller than the standard estimate of N based on levels of variability and an estimate of the mutation rate, which was interpreted as a genome-wide excess of linkage disequilibrium in the two species. Our study presents the advantages of consistently sampled data and more accurate estimates of ρ.

METHODS

A list of definitions for the symbols used in this paper (in approximate order of introduction) can be found in Table 1.

Loci and samples: We consider 29 loci collected from a single *D. simulans* population (Wolfskill Orchard, California), reported by Begun and Whitley (2000). Sequences were obtained from GenBank and aligned by eye. We consider only biallelic single-nucleotide base substitutions; multiple substitutions, insertion-deletion mutations, and other overlapping mutational events (*e.g.*, substitutions overlapping with deletions) are excluded from analyses. Excluding multiple substitutions should not lead to a bias in our frequency spectrum analyses and should be conservative for our estimates of ρ. The alternative, *i.e.*, considering multiple substitutions as multiple mutations with missing information, would lead to underestimates of H and R_{M} due to the missing information and thus to underestimates of ρ (see below).

Estimating r: Previous methods for estimating r have fit high-order polynomial curves to the available genetic and physical map data for *D. simulans* (True *et al.* 1996; Andolfatto and Przeworski 2000). Here, we assume a composite model, where the rate is constant across much of the chromosome, but is reduced near the telomeres and centromeres (*cf.* Charlesworth 1996; Andolfatto and Przeworski 2001). Cytological map positions are obtained from Flybase (http://flybase.bio.indiana.edu; Dec 1, 2000), and the DNA content for each band is estimated from the *D. melanogaster* values in Heino *et al.* (1994; *i.e.*, we assume *D. melanogaster* and *D. simulans* have similar amounts of DNA per band). There is no recombination in males, so we use sex-averaged rates.

X chromosome: Genetic map positions for 14 marker loci are taken from Sturtevant (1929). Note that these marker loci and the autosomal ones are distinct from the loci considered in the polymorphism analysis. The X chromosome is divided into telomeric (I), middle (II), and centromeric (III) segments (see Figure 1a). We model genetic distance as a linear function of physical distance from the *white* locus (2.9 Mb, 4.1 cM) to the *fused* locus (20.2 Mb, 59.4 cM; region II, Pearson $R^2 >$ 0.99). The recombination rate estimate over this interval is 3.2 cM/Mb. Recent recombination measurements (Takano-Shimizu 1999) have estimated lower rates of recombination near the telomere (*i.e.*, region I) than measured by Sturtevant (1929). Combining the ge-

TABLE 1

Definitions of symbols

Symbol	Meaning
D	Measure of the frequency spectrum from Tajima (1989a).
N	Diploid effective population size.
r	Sex-averaged recombination rate per base pair per generation.
ρ	$4Nr$.
$\hat{\rho}$	Estimate of ρ from sequence data (C_{HRM}, *cf.* Wall 2000).
$\hat{r}$	Laboratory-based estimate of r (see methods).
$\hat{N}_\rho$	$\hat{\rho}/4\hat{r}$.
H	Number of distinct haplotypes.
R_{M}	Minimum number of inferred recombination events (*cf.* Hudson and Kaplan 1985).
N_{x}	One-half of the haploid effective population size for the X chromosome.
N_{a}	Diploid (*i.e.*, one-half of the haploid) effective population size for the autosomes.
$\hat{N}_{\rho\text{x}}$	Estimate of N_{x} for the actual data using H, R_{M}, and $\hat{r}$ (see methods).
$\hat{N}_{\rho\text{a}}$	Estimate of N_{a} for the actual data using H, R_{M}, and $\hat{r}$ (see methods).
μ	Sex-averaged mutation rate per base pair per generation.
$\hat{\mu}$	Estimate of μ of 1.5×10^{-9} per base pair per generation (see methods).
θ	$4N\mu$.
S	The observed number of segregating sites.
$(1 - \delta_\theta)$	The decrease in diversity due to hitchhiking (or bottlenecks).
Λ_r	The rate of selective sweeps (*cf.* Braverman *et al.* 1995).
T_0	Time (scaled in units of $2N$ generations) of the bottleneck.
N_{b}	Effective population size at time T_0.
n_r	Ratio of $N_{\text{x}}/N_{\text{a}}$ used for diversity estimates and timescaling of bottleneck simulations.
$\overline{D}$	Average D value (for all X-linked loci or all autosomal loci).
R	Likelihood-ratio statistic (see methods).
$\hat{N}_{\rho\text{x}0}$	Maximum-likelihood estimate of N_{x} from H, R_{M}, and $\hat{r}$ for simulated data with $N_{\text{x}} = N_0$ (see methods).
$\hat{N}_{\rho\text{a}1}$	Maximum-likelihood estimate of N_{a} from H, R_{M}, and $\hat{r}$ for simulated data with $N_{\text{a}} = N_1$ (see methods).
R^*	Likelihood that $\hat{N}_{\text{x}}$ and $\hat{N}_{\text{a}}$ from simulated data equal the actual population size estimates, assuming the null model.
$\hat{\theta}_{\text{W}}$	Estimate of θ from Watterson (1975).
$\hat{N}_{\theta\text{a}}$	Estimate of N from $\hat{\theta}_{\text{W}}$ (for the autosomes) and $\hat{\mu}$.
$\hat{N}_{\theta\text{x}}$	Estimate of N from $\hat{\theta}_{\text{W}}$ (for the X) and $\hat{\mu}$.

netic map data of Sturtevant (1929) and Takano-Shimizu (1999), and assuming that genetic distance increases with the square of physical distance from the telomere (0.0 Mb) to the *white* locus (2.9 Mb), a rate of 1.5 cM/Mb is estimated for *Pgd* (2.1 Mb). All other X-linked loci from Begun and Whitley (2000) fall within the boundaries of region II.

Chromosome 3R: Genetic map positions for 13 marker loci were obtained from Ohnishi and Voelker (1979). *D. simulans* and *D. melanogaster* differ by a single inversion on this chromosomal arm (breakpoints 84F1 and 93F6-7; Lemeunier and Ashburner 1976). This difference has been accounted for in estimates of physical distances between genetic markers. The chromosome arm is divided into two regions (see Figure 1b). The first region is defined as the centromeric region (I) and extends from the centromere to *delta* (5.4 Mb, 64 cM). We define a second region (II, Figure 1b) that encompasses the rest of the chromosome arm from *delta* to *Acph-1* (22.2 Mb, 134 cM), which is an estimated 16.8 Mb in length. For region II, we model genetic distance as a linear function of physical distance (Pearson $R^2 >$ 0.99). Under this model, the recombination rate is estimated to be 4.2 cM/Mb.

If we assume that genetic distance increases with the square of physical distance from the centromere (0.0 Mb) to *delta* (5.4 Mb), we estimate a rate of 2.2 cM/Mb for the *miranda* locus (4.9 Mb). We localized the *pitchoune* locus (93F16-94A1) outside of the distal breakpoint (93F6-7) of the fixed 3R inversion difference between *D. melanogaster* and *D. simulans* by *in situ* hybridization on polytene chromosomes (using a method modified from Sniegowski and Charlesworth 1994). Since *pitchoune* lies outside the inversion, it resides ~15.5 Mb from the centromere (*i.e.*, region II). The remaining 12 loci on chromosome 3R considered for polymorphism analyses are also in region II.

Estimating N_ρ: We proceed assuming that the true r is known for each locus. Two summaries of linkage disequilibrium are H, the observed number of distinct haplotypes, and R_{M}, the minimum number of inferred recombination events (Hudson and Kaplan 1985). We calculate $\Pr(H, R_{\text{M}}|\rho)$, or equivalently

$$\Pr(H, R_{\text{M}}|N) \propto \text{lik}(N|H, R_{\text{M}}),$$

assuming the null model. Likelihoods are calculated for each locus separately, as well as for all X-linked loci and all autosomal loci (see below). Here "lik" refers to the

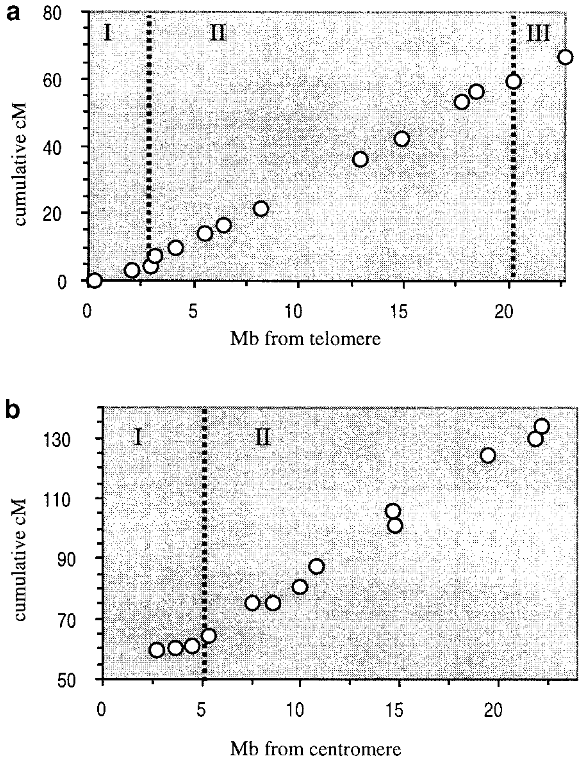

Figure 1.—Cumulative genetic distance *vs.* cumulative physical distance based on 14 genetic markers for the X chromosome (a) and 13 genetic markers for chromosome 3R (b). The X chromosome is divided into telomeric (I) and middle (II) and centromeric (III) segments. The *white* locus (2.9 Mb) and *fused* locus (20.2 Mb) are chosen as boundaries (dotted lines). Chromosome 3R is divided into centromeric (I) and middle (II) segments. The *delta* locus (5.4 Mb) defines the boundary between these two regions (dotted line).

likelihood, and the joint likelihood at a collection of loci is found by multiplying the likelihoods at the individual loci. This is reasonable since none of the loci are closely linked to each other and thus can be considered as evolutionarily independent. Define N_x and N_a as the effective population sizes of the X and the autosomes, respectively. Define $\hat{N}_{\rho x}$ and $\hat{N}_{\rho a}$ as the maximum-likelihood estimates for the X-linked and autosomal loci; $\hat{N}_{\rho x}$ is the value of N_x that maximizes $\Pi_{\text{all X-linked loci}}\text{lik}(N_x|H, R_M)$, while $\hat{N}_{\rho a}$ is the value of N_a that maximizes $\Pi_{\text{all autosomal loci}}\text{lik}(N_a|H, R_M)$. We choose to summarize the data before performing maximum likelihood because of computational constraints; maximum likelihood on the full data is computationally infeasible even for recombination rates one-tenth of those considered here. As it is, the likelihoods for this article took several months of computing time on a pair of 600 Mhz Pentium III processors.

The likelihoods were estimated from simulations that assume a neutral infinite-sites model and use the protocol of Hudson (1993). This method differs slightly from standard coalescent simulations, which generate random genealogies and then place mutations on the branches with rate $\theta/2$. ($\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per base pair per generation.) Instead, we generate random genealogies and then place the observed number of mutations, S, on the tree. One motivation for this procedure is that S is observed, while θ must be estimated (Hudson 1993). Both methods produce similar results in other contexts (*e.g.*, Wall 2000; Wall and Hudson 2001). A minimum of 2×10^5 replicates were run for a large number of different N_x and N_a values at each locus. The particular values used and estimated likelihoods are available on request from the authors.

Null simulations: We compare the selective sweep and bottleneck models (described below) with a constant-sized, panmictic, neutral coalescent model (Hudson 1983). Here, we consider it more appropriate to use standard coalescent simulations (with fixed mutation rate, as opposed to a fixed number of segregating sites) for the selective sweep simulations, because they have a large variance in tree sizes. To ensure comparability, we also use standard coalescent simulations for the bottleneck and null simulations. To estimate θ, the population mutation rate, we split the data into three classes of sites (introns, synonymous sites, and nonsynonymous sites) and assume a fixed neutral (population) mutation rate for each class. θ is then estimated for each class using Watterson's (1975) estimator, separately for the X and 3R.

Selective sweep simulations: We consider a model where recurrent, nonoverlapping favorable alleles arise at sites linked to a neutral locus. The model assumes that beneficial mutations are selected immediately upon introduction into the population. Our methods follow those of Braverman *et al.* (1995), but we run coalescent simulations with a fixed mutation rate as opposed to the "fixed S" methodology (Hudson 1993; Braverman *et al.* 1995). Also, our implementation incorporates intragenic recombination within the neutral locus, except during the selective phases. The lack of intragenic recombination during the selective phases makes little difference to our results, so long as the neutral locus is relatively short (M. Przeworski, unpublished results). On average, it produces more linkage disequilibrium than a model that always has intragenic recombination (thus is conservative for our purposes). Note that there is a typo in Equation 1 of Braverman *et al.* (1995), describing the increase in frequency of the favored allele. Instead, we use Equation 3a in Stephan *et al.* (1992), which approximates the increase from frequency ε to $1 - \varepsilon$. We implement our simulations with $\varepsilon = 1/2N$ (as in Braverman *et al.* 1995). Selection is additive, and we arbitrarily assume the selection coefficient of one beneficial allele to be $s = 0.005$. We obtain similar results for other values of s (results not shown; see also discussion).

Denote the decrease in diversity due to hitchhiking by $(1 - \delta_\theta)$. We choose three plausible values for δ_θ: 0.85, 0.75, and 0.65 for the autosomes and 0.65, 0.55, and 0.45 for the X. The autosomal values were chosen to be close to the observed ratio of non-African to African diversity levels (Andolfatto 2001), while the X chromosome values were chosen to produce a wide range of X-linked to autosomal diversity ratios. The motivation here is that African populations of *D. simulans* may be roughly at equilibrium, whereas non-African populations might have reduced variation due to local adaptation. This line of reasoning suggests a range of plausible values for δ_θ. Note also that data from other loci suggest that the ratio of X to autosomal levels of diversity may be higher than was observed by Begun and Whitley (2000; 0.79 for all non-African data, *cf.* Andolfatto 2001). θ is estimated for each locus as in the null simulations and then divided by δ_θ. The rate of selective sweeps (Λ_r in Braverman *et al.* 1995) was estimated by simulation to produce the desired decrease $(1 - \delta_\theta)$ in θ. The actual values used are listed in Table 2. In almost all of the 3R simulations and for larger values of N_x, Λ_r is small enough that the probability of a second benefical mutation arising while a sweep is still ongoing is <0.05. See the discussion for more on the applicability of the model.

Bottleneck simulations: It is relatively straightforward to incorporate changes in population size into the coalescent framework (*e.g.*, Tajima 1989b; Slatkin and Hudson 1991). For the autosomes, we consider a model where the effective population size is constant in the past at 5×10^6. Then, at time T_0, the population size immediately crashes to a bottleneck size N_b, after which it increases exponentially to a current effective population size of 5×10^6. For a given value of T_0, we choose N_b so that the reduction in diversity caused by a bottleneck equals the autosomal values of $(1 - \delta_\theta)$ (see above). As before, we assume a fixed mutation rate for introns, synonymous sites, and nonsynonymous sites, and we estimate these rates from the data (*cf.* Watterson 1975) before dividing by δ_θ. Since one of our goals is to examine whether bottlenecks have a stronger effect on the X, we use the estimated autosomal mutation rates for the X chromosome as well, after multiplying by n_r to correct for the chromosomal differences in effective population size. The scaled time T_0 (in coalescent time units) is similarly divided by n_r for the X relative to 3R, but the population sizes estimated from the patterns of linkage disequilibrium (described below) are assumed to freely vary. This need arises because θ values for X-linked loci must be close to Watterson's (1975) estimate of θ for simulations to be comparable (*i.e.*, levels of diversity in the simulations should be close to what is observed in the data). However, it is better to allow N_x and N_a (as estimates of linkage disequilibrium) to vary freely, so that we can see what effect bottlenecks have on linkage disequilibrium. We present results for the following parameter combinations: (a) $T_0 = 2000$ generations ago, $\delta_\theta = 0.85$, and $n_r = 0.6$; (b) $T_0 = 2000$ generations ago, $\delta_\theta = 0.75$, and $n_r = 0.7$; (c) $T_0 = 1.2 \times$

TABLE 2

Rate of selective sweeps for hitchhiking simulations

X						3R					
$\delta_\theta = 0.45$[a]		$\delta_\theta = 0.55$[b]		$\delta_\theta = 0.65$		$\delta_\theta = 0.65$		$\delta_\theta = 0.75$[b]		$\delta_\theta = 0.85$[a]	
N ($\times 10^6$)	Λ_r	N ($\times 10^6$)	Λ_r	N ($\times 10^6$)	Λ_r	N ($\times 10^6$)	Λ_r	N ($\times 10^6$)	Λ_r	N ($\times 10^6$)	Λ_r
0.04	0.00044	0.03	0.000194	0.03	0.000104	0.1	0.000134	0.1	0.000078	0.1	0.00004
0.08	0.000376	0.06	0.000190	0.06	0.000104	0.2	0.00014	0.2	0.000082	0.2	0.000042
0.12	0.000330	0.09	0.000186	0.09	0.000104	0.3	0.000144	0.3	0.000086	0.3	0.000044
0.16	0.000298	0.12	0.000182	0.12	0.000104	0.4	0.00015	0.4	0.00009	0.4	0.000046
0.20	0.000282	0.15	0.000178	0.15	0.000106	0.5	0.000154	0.5	0.000094	0.5	0.000048
0.24	0.000282	0.18	0.000174	0.18	0.000106	0.6	0.000156	0.6	0.000094	0.6	0.00005
0.28	0.000282	0.21	0.00017	0.21	0.000106	0.7	0.000158	0.7	0.000096	0.7	0.00005
0.32	0.000282	0.24	0.00017	0.24	0.000106	0.8	0.00016	0.8	0.000096	0.8	0.000052
0.36	0.000282	0.27	0.000172	0.27	0.000108	0.9	0.000162	0.9	0.000098	0.9	0.000052
0.40	0.000282	0.30	0.000174			1.0	0.000166	1.0	0.000098		
0.44	0.000282	0.33	0.000174			1.1	0.000168				

[a] Corresponds to Figure 4a.
[b] Corresponds to Figure 4b.

TABLE 3

Parameter values used in bottleneck simulations

				Autosomes		X	
	δ_θ	n_r	$\hat{N}_{\theta x}/\hat{N}_{\theta a}$[a]	T_0[b]	N_b	T_0[b]	N_b
(a)[c]	0.85	0.6	0.538	1×10^{-4}	909	1.667×10^{-4}	545
(b)[d]	0.75	0.7	0.617	1×10^{-4}	455	1.429×10^{-4}	318
(c)[e]	0.75	0.7	0.620	6×10^{-3}	50,000	8.571×10^{-3}	35,000

See the text and Table 1 for parameter definitions.

[a] The observed ratio of X to autosomal diversity levels in the simulations.

[b] Scaled in units of $4N$ generations.

[c] Same parameter values as Figure 5a.

[d] Same parameter values as Figure 5b.

[e] Same parameter values as Figure 5c.

10^5 generations ago, $\delta_\theta = 0.85$, and $n_r = 0.7$. T_0 values were chosen to correspond to recent (∼200 years ago) or ancient (∼12,000 years ago) colonization of the Americas, δ_θ was chosen to be close to the ratio of non-African to African autosomal diversity levels in *D. simulans* (0.76, *cf.* Table 3 in Andolfatto 2001), and n_r was chosen so that the relative levels of variability on the X and the autosomes would be close to what was observed by Begun and Whitley (2000). These values are shown in Table 3. See the discussion for more on the sensitivity of the results to the particular parameter values chosen and whether the particular n_r values used are plausible.

Frequency spectrum of segregating mutations: We use D (Tajima 1989a) to test whether the observed frequency spectrum of segregating mutations is compatible with the expectations under bottlenecks or selective sweeps. We consider $\overline{D}$, the average D value for the X-linked loci and the 3R loci, separately and tabulate both the average simulated $\overline{D}$ value and the proportion of simulations that have $\overline{D}$ greater than or equal to what is observed (see results). A total of 5000 replicates were run for each model and parameter combination. We present results for only the most conservative values of N_x and N_a.

Estimating μ: We take a value of 1.5×10^{-9} per site per generation for the neutral mutation rate at silent sites. This estimate is based on average per year divergence at synonymous sites in various Drosophila species comparisons (Sharp and Li 1989; Li 1997; McVean and Vieira 2001), assuming an average of 10 generations per year for *D. simulans* (see Andolfatto and Przeworski 2000 for a discussion). The neutral mutation rate is only loosely related to our analyses, but it provides a connection between observed levels of diversity and an estimate of the effective population size under the null model. We assume that mutation rates do not vary significantly among chromosomes. This assumption is supported by comparisons of average divergence at synonymous sites on the X and on chromosome 3R (Begun and Whitley 2000).

Credibility intervals for $N_{\theta x}/N_{\theta a}$: To assess what range of X to autosomal diversity levels is consistent with the data of Begun and Whitley (2000), we calculate approximate credibility intervals for $N_{\theta x}/N_{\theta a}$ from the observed numbers of segregating sites. We take a neutral mutation rate of $\hat{\mu} = 1.5 \times 10^{-9}$ per site per generation (see above). As with the bottleneck simulations, we assume fixed population mutation rates for synonymous sites, nonsynonymous sites, and introns and estimate these (*cf.* Watterson 1975) from the autosomal loci. We then assume that these parameters (multiplied by $N_{\theta x}/N_{\theta a}$) apply to the X-linked loci as well and calculate

$$\mathrm{lik}(N_{\theta x}/N_{\theta a}|S = \text{observed value}) \propto \Pr(S = \text{observed value}|N_{\theta x}/N_{\theta a}).$$

Here S refers to the total number of inferred segregating sites summed over all X-linked loci. We employ the standard χ^2 approximation for $-2\ \ln(L_1/L_0)$ to obtain approximate 95% credibility intervals, where L_0 is the maximum likelihood and L_1 is the likelihood at an alternative parameter value.

Likelihood-based statistics: To quantify how consistent the actual data are with the null model, we employ a likelihood-ratio test. We calculate $\hat{N}_{\rho x}$ and $\hat{N}_{\rho a}$ from the actual data as described earlier. Then, for $N_x = N_0$ and $N_a = N_1$, we calculate

$$R(N_0, N_1) = \left(\prod_{\text{all X-linked loci}} \frac{\mathrm{lik}(\hat{N}_{\rho x}|H, R_M)}{\mathrm{lik}(N_0|H, R_M)}\right)\left(\prod_{\text{all 3R loci}} \frac{\mathrm{lik}(\hat{N}_{\rho a}|H, R_M)}{\mathrm{lik}(N_1|H, R_M)}\right).$$

The significance levels for R are determined by simulation for a range of N_0 and N_1 values. We simulate 10^4 replicates of the 29 loci with $N_x = N_0$ and $N_a = N_1$. Then, we calculate $\hat{N}_{\rho x0}$ and $\hat{N}_{\rho a1}$ for each replicate, where $\hat{N}_{\rho x0}$ is the value of N_x that maximizes

$$\prod_{\substack{\text{all simulated X-linked loci}\\ \text{in a replicate with } N_X = N_0}} \mathrm{lik}(N_x|H, R_M)$$

and $N_{\rho a1}$ is the value of N_a that maximizes

$$\prod_{\substack{\text{all simulated autosomal loci}\\ \text{in a replicate with } N_a = N_1}} \mathrm{lik}(N_a|H, R_M).$$

The collection of

$$\left(\prod_{\substack{\text{all simulated X-linked loci}\\ \text{in a replicate with } N_x=N_0}} \frac{\text{lik}(\hat{N}_{\rho x0}|H,\ R_M)}{\text{lik}(N_0|H,\ R_M)}\right)$$
$$\left(\prod_{\substack{\text{all simulated 3R loci}\\ \text{in a replicate with } N_a=N_1}} \frac{\text{lik}(\hat{N}_{\rho a1}|H,\ R_M)}{\text{lik}(N_1|H,\ R_M)}\right)$$

values provides a simulated distribution of $R(N_0,\ N_1)$ values, from which we tabulate how often the simulated R values are greater than or equal to the actual R value. For each parameter combination, we also calculate what proportion of trials have estimated population sizes $\hat{N}_{\rho x0}$ and $\hat{N}_{\rho a1}$ equal to the values estimated from the actual data. Define

$$R^*(N_0,\ N_1) = \Pr(\hat{N}_{\rho x0} = \hat{N}_{\rho x}|N_x = N_0)$$
$$\times \Pr(\hat{N}_{\rho a1} = \hat{N}_{\rho a}|N_a = N_1).$$

$R^*(N_0,\ N_1)$ is the likelihood of the actual effective population size estimates (using H and R_M) when data are generated under the null model (with $N_x = N_0$ and $N_a = N_1$). From our simulations, we plot the value of R^* as a function of N_0 and N_1.

Ideally, we would like to perform the same analyses under the selective sweep and bottleneck models, but calculating the relevant likelihoods is computationally prohibitive. Instead, we use R^* again, with all likelihoods calculated assuming the null model, even though the simulated data are generated under a different model. As before, R^* is a measure of how likely it is for the simulated data to produce the actual estimated population sizes. A total of 10^4 replicates are run for each parameter combination. Other *ad hoc* statistics were considered; they all produced similar results (results not shown).

RESULTS

Excess linkage disequilibrium on all chromosomes: Table 4 shows the estimates of $\hat{N}_\rho$ on the basis of H, R_M, and $\hat{r}$ for each locus. One observation that is immediately apparent is that these values are systematically less than estimates of N based on estimates of the neutral mutation rate and observed levels of polymorphism. For example, if we take $\hat{\mu} = 1.5 \times 10^{-9}$/bp/generation for silent sites (see METHODS) and $\theta = 0.030$ per synonymous base pair (estimated from all of the autosomal loci considered in this article, *cf.* WATTERSON 1975), then we obtain the estimate $\hat{N}_{\theta a} = 5.0 \times 10^6$. The corresponding estimate from the X-linked loci is $\hat{N}_{\theta x} = 2.5 \times 10^6$. In contrast, 26 out of 29 loci have $\hat{N}_\rho$ estimates at least an order of magnitude less than the corresponding $\hat{N}_\theta$ estimate. This discrepancy between $\hat{N}_\rho$ (estimated from linkage disequilibrium) and $\hat{N}_\theta$ (estimated from levels of diversity) has been noted before with different Drosophila data and slightly different methodology (ANDOLFATTO and PRZEWORSKI 2000) and implies that there is more linkage disequilibrium in nucleotide polymorphism data than expected from standard estimates of μ, r, and θ. Note that the low values in Table 4 are not the result of the particular properties of $\hat{\rho} = C_{\text{HRM}}$. In fact, simulations (under the standard equilibrium neutral model) show that for the small sample sizes considered here, C_{HRM} is biased upward, suggesting that the $\hat{N}_\rho$ values in Table 4 might on average be overestimates (J. D. WALL, unpublished results).

TABLE 4

$\hat{N}_\rho$ estimates for each locus

X		3R	
Locus	$\hat{N}_{\rho x}$	Locus	$\hat{N}_{\rho a}$
pgd	0	*mir*	3.5×10^6
mei-9	0	*nos*	1.7×10^5
ovo	0	*osa*	9.3×10^4
X	6.3×10^4	*hsc70*	0
sqh	0	*cp190*	0
ct	1.2×10^4	*hyd*	7.3×10^4
dec-1	7.7×10^4	*rel*	1.6×10^5
sn	1.3×10^5	*pit*	5.0×10^5
otu	1.4×10^5	*ap50*	3.8×10^5
yp3	0	*tcp1*	4.1×10^5
gar	0	*fzo*	3.2×10^5
sog	0	*aats*	2.4×10^5
rud	0	*tld*	3.0×10^6
bnb	2.5×10^5	*osbp*	7.6×10^5
mei-218	0		

For comparison, the estimates of N from levels of diversity are $\hat{N}_{\theta x} = 2.5 \times 10^6$ and $\hat{N}_{\theta a} = 5.0 \times 10^6$.

Contrasting patterns between X and autosomes: Because patterns of variation vary greatly from locus to locus even when the underlying parameters are the same, the precision of the estimate of N can be greatly increased by combining information from multiple loci. Figure 2 shows the relative log likelihoods of N for all of the X-linked loci (the curve on the left) and all of the autosomal loci (the curve on the right). For ease of comparison, the curves have been normalized so that their maxima are at 0. It is striking how distinct the two likelihood curves are: $\hat{N}_{\rho a}$ ($= 3.2 \times 10^5$) is more than six times $\hat{N}_{\rho x}$ ($= 0.5 \times 10^5$). The horizontal line in Figure 2 shows the ~95% credibility intervals (using the standard asymptotic approximations for maximum likelihood) for the chromosome-specific estimates of N; the two intervals do not overlap. A nonparametric rank order test shows that the locus-specific $\hat{N}_\rho$ estimates for the X-linked loci are indeed less than the autosomal estimates (Table 4, Mann-Whitney U-test; $P < 0.002$).

Note that since males carry only one X chromosome, we do not necessarily expect N_x to equal N_a. If male and female effective population sizes are equal, then $4N_x = 3N_a$. However, there are many possible factors that may cause the two effective population sizes to be unequal (CROW and MORTON 1954; CABALLERO 1995; CHARLES-

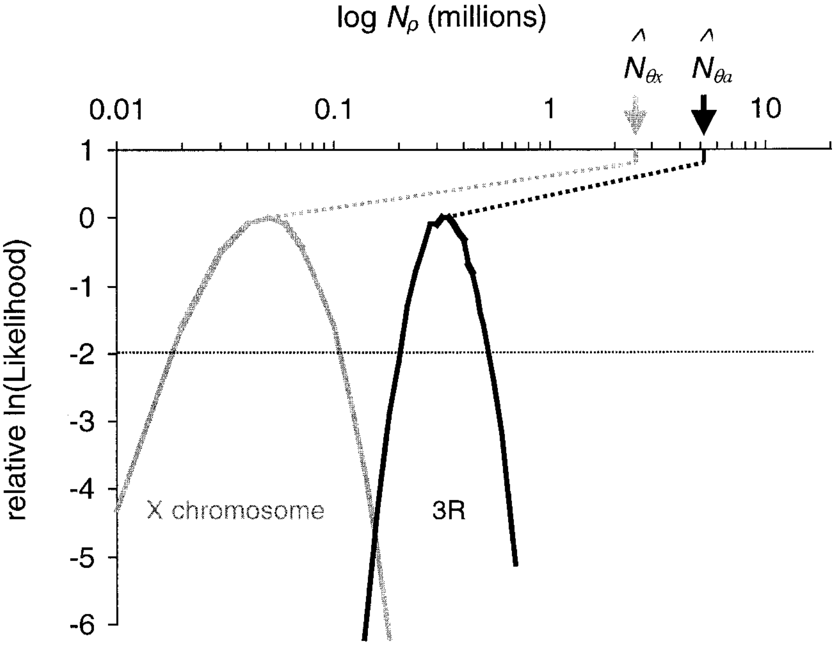

FIGURE 2.—Plots of the relative values of ln(lik($N|H$, R_M)) as a function of N_ρ for the X-linked loci (curve on the left) and the 3R loci (curve on the right). Both have been normalized so that their maxima are at 0. Corresponding estimates of $\hat{N}_\theta$ are indicated by arrows. The horizontal line is at −1.92, which delineates the ~95% credibility intervals (see METHODS). A minimum of 2×10^5 replicates were run for each locus and each value of N. The values of N_ρ considered were 1.0×10^4–8.0×10^4 (increment 1.0×10^4) and 1.0×10^5–7.0×10^5 (increment 2.0×10^4).

WORTH 2001). If chromosomal effective population sizes are proportional to silent site $\hat{\theta}_W$ values (*cf.* WATTERSON 1975), then, from the data analyzed here, $\hat{N}_{\theta x}/\hat{N}_{\theta a} \approx 0.50$. When all sites are considered (with different rates for synonymous sites, nonsynonymous sites, and introns), then $\hat{N}_{\theta x}/\hat{N}_{\theta a} \approx 0.59$. The ~95% credibility interval for $N_{\theta x}/N_{\theta a}$ based on all sites (see METHODS) is 0.43–0.69. Under neutrality, $N_{\theta x}/N_{\theta a} = 0.50$ is unexpected, regardless of how biased the gender-specific population sizes are (CABALLERO 1995). This observation is in part what led BEGUN and WHITLEY (2000) to conclude that natural selection must be acting to reduce the levels of variation on the X relative to the autosomes. Our results demonstrate that the difference in the levels of linkage disequilibrium between the X and 3R ($\hat{N}_{\rho x}/\hat{N}_{\rho a} = 0.16$) is substantially greater than the difference in their diversity levels.

Figure 3a shows the P value of R (see METHODS) as a function of N_x and N_a. For all population sizes where $2N_x \geq N_a$, the actual value of R is significantly too large. This suggests that there is significantly more linkage disequilibrium on the X than on 3R, even after correcting for the differences in effective population sizes suggested by diversity levels. For the same population sizes, Figure 3b shows the proportion of trials for which $\hat{N}_{\rho x0} = 0.5 \times 10^5$ and $\hat{N}_{\rho a1} = 3.2 \times 10^5$ (see METHODS). The different shading categories were chosen so that in Figure 3, a and b were as similar in appearance as possible; the lightest areas on both graphs represent areas of parameter space that are compatible with the data. For all population sizes where $2N_x \geq N_a$, the value of R^* is quite small (*i.e.*, $R^* < 8.0 \times 10^{-4}$). If instead we repeat the rank order test with the null hypothesis that $2N_x = N_a$, then the two chromosomes are still significantly different (Table 4, Mann-Whitney U-test; $P < 0.01$).

The chromosomal difference in diversity levels (BEGUN and WHITLEY 2000), the chromosomal difference in levels of linkage disequilibrium (this study), and the overall high levels of linkage disequilibrium (ANDOLFATTO and PRZEWORSKI 2000; this study) are all ways in which *D. simulans* data do not conform to the expectations of the standard equilibrium neutral model. We now examine how two simple alternative models (a recurrent selective sweep model and a recent bottleneck model) are expected to affect the levels of diversity and linkage disequilibrium on different chromosomes. Though the true history of *D. simulans* populations is likely to be much more complex, these models should help us gain insight into the effects of demography and natural selection on the patterns of segregating variation.

Sweep model: We considered all combinations of $\delta_\theta = 0.85$, 0.75, and 0.65 for 3R and $\delta_\theta = 0.65$, 0.55, and 0.45 for the X. All nine sets of simulations produced very similar results, and we display only a representative pair of them here. Figure 4 shows the value of R^* as a function of N_x and N_a. Figure 4a has $\delta_\theta = 0.85$ for 3R and $\delta_\theta = 0.45$ for the X, while the corresponding δ_θ values in Figure 4b are 0.75 and 0.55, respectively. We find that recurrent selective sweeps do not lead to striking increases in levels of linkage disequilibrium, as measured (see also PRZEWORSKI 2002). In particular, the increase in average $\hat{N}_{\rho x0}$ and $\hat{N}_{\rho a1}$ in the sweep simulations is no more than what is expected from the decrease in levels of diversity. In other words, the estimated ratio of the number of recombination events to the number of mutation events, $\hat{\rho}/\hat{\theta}_W$, does not vary much when data are generated under either the null model or the recurrent selective sweep model. Exploratory simulations suggest that this observation might hold under a

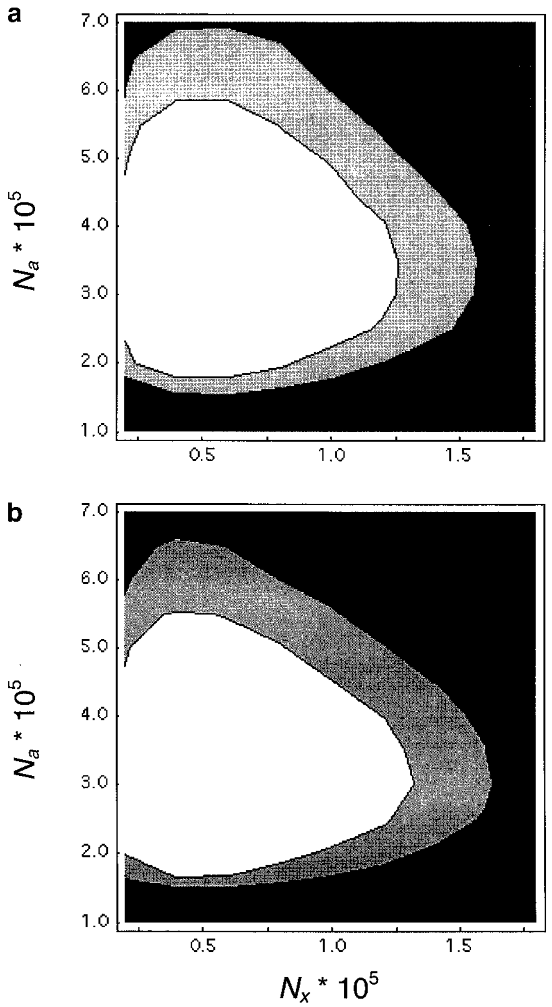

FIGURE 3.—Unusualness of the data as a function of N_x and N_a. a plots the P value of R, while b plots the value of R^*. (□) $P > 0.05$; (▨) $0.01 < P < 0.05$; (■) $P < 0.01$. See METHODS for details. The cutoffs for the different shading categories in b were chosen so that the appearances of the two figures were as similar as possible. (□) $R^* > 10^{-3}$; (▨) $2 \times 10^{-4} < R^* < 10^{-3}$; (■) $R^* < 2 \times 10^{-4}$. For comparison, $\hat{N}_{\theta x} = 25 \times 10^5$ and $\hat{N}_{\theta a} = 50 \times 10^5$.

wider range of sample sizes and relative recombination rates than considered for the *D. simulans* data (results not shown). Thus, this simple model for repeated episodes of positive selection seems to explain neither the overall high levels of linkage disequilibrium nor the chromosomal difference in levels of linkage disequilibrium.

Previous work has shown that recurrent selective sweeps lead to a strong skew in the frequency spectrum toward an excess of rare variants (BRAVERMAN *et al.* 1995). The *D. simulans* data of BEGUN and WHITLEY (2000), on the other hand, show no marked skew in the frequency spectrum. The average value of Tajima's D for the 15 X-linked loci is 0.205, while the average value for the 14 loci on 3R is −0.021. Table 5 shows what proportions of the simulations have average D values

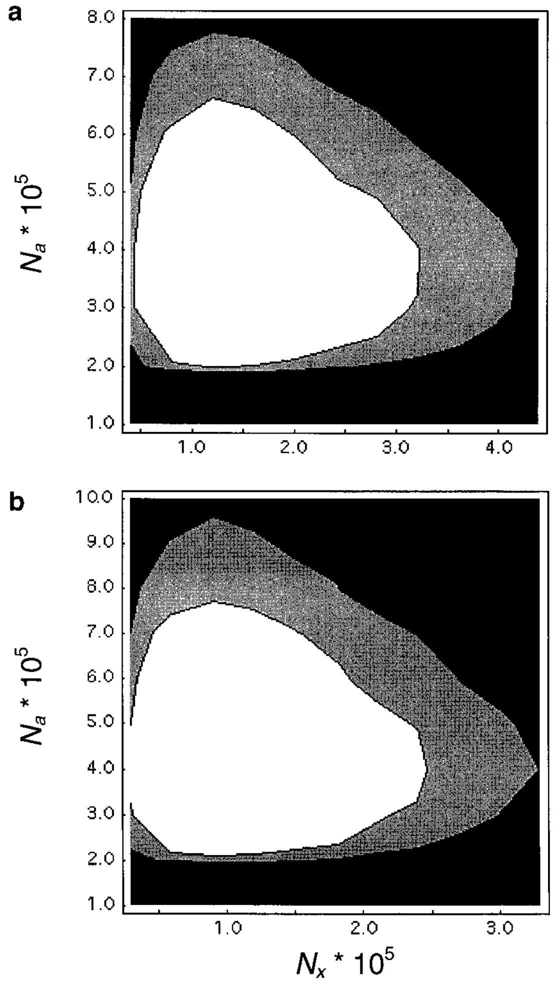

FIGURE 4.—R^* as a function of N_x and N_a under a model of recurrent selective sweeps. a has $\delta_\theta = 0.85$ for 3R and $\delta_\theta = 0.45$ for the X, while b has $\delta_\theta = 0.75$ for 3R and $\delta_\theta = 0.55$ for the X. See METHODS for more details. The shading categories are the same as in Figure 3b.

($\bar{D}$) greater than or equal to what is actually observed. Under recurrent hitchhiking, the average simulated $\bar{D}$ is negative, as expected. The actual $\bar{D}$ for the X-linked loci is significantly too high ($P < 0.004$, one-tailed test), while the true $\bar{D}$ for the autosomal loci is not unusual.

Bottleneck model: Due to computational constraints, we consider only a few parameter combinations. Figure 5 shows R^* as a function of N_x and N_a for three different examples. As can be seen, recent bottlenecks are consistent with much higher effective population sizes. Equivalently, recent bottlenecks cause an increase in observed levels of linkage disequilibrium. For example, two out of three (Figure 5, a and b) are consistent with $N_a = 5.0 \times 10^6$, while one out of three (Figure 5b) is consistent with $N_x = 2.5 \times 10^6$. In addition, at least for some parameter values (*e.g.*, those of Figure 5b), the average ratio of $\hat{N}_{\rho x0}/\hat{N}_{\rho a1}$ is much larger than what was estimated under the null model (0.16) and closer to the expectation from levels of diversity (*i.e.*, 0.59). Since there are

TABLE 5

Frequency spectrum test for various models

X			3R		
Model	Avg. $\overline{D}$	Pr($\overline{D} \geq 0.205$)	Model	Avg. $\overline{D}$	Pr($\overline{D} \geq -0.021$)
Null	−0.049	0.1296	Null	−0.034	0.4712
Sweep, $\delta_\theta = 0.45$[a]	−0.416	$\underline{0.0030}$	Sweep, $\delta_\theta = 0.85$[a]	−0.113	0.2798
Sweep, $\delta_\theta = 0.55$[b]	−0.354	$\underline{0.0036}$	Sweep, $\delta_\theta = 0.75$[b]	−0.190	0.1572
Bottleneck (a)	0.317	0.6794	Bottleneck (a)	0.238	0.9118
Bottleneck (b)	0.405	0.7708	Bottleneck (b)	0.351	0.9706
Bottleneck (c)	0.347	0.7188	Bottleneck (c)	0.319	0.9500

Underlined values have $P < 0.01$. See the text and Tables 2 and 3 for the details of the models.

[a] Same parameter values as in Figure 4a.

[b] Same parameter values as in Figure 4b.

fewer X chromosomes than autosomes, a bottleneck is more severe for the X (*i.e.*, the minimal population size is smaller). This leads to both a greater increase in linkage disequilibrium and a greater reduction in levels of variability on the X relative to the autosomes. In principle, a recent bottleneck might explain both the chromosomal differences in levels of linkage disequilibrium and the overall high levels of linkage disequilibrium, but it remains to be seen whether the parameter values required are plausible (see DISCUSSION).

The effect of a bottleneck on the frequency spectrum is complex. For results from a similar model, see Fay and Wu (1999). During and immediately after a bottleneck, a deficiency of rare variants is expected, but as time passes, the accumulation of recent mutations leads to an excess of low frequency segregating sites. For the small values of T_0 considered here, bottlenecks are expected to lead to positive Tajima's D values (more so for the X than the autosomes). Table 5 shows the average of the simulated $\overline{D}$ values, as well as the proportion of simulated $\overline{D}$ values greater than or equal to the actual values. As is expected under a recent bottleneck, the actual $\overline{D}$ is higher for the X-linked loci than it is for the autosomal loci. In all cases, the actual $\overline{D}$ values for both the X and 3R are within the middle 95% of the simulated distribution, though $\overline{D}$ for the 3R loci is close to being significantly too low.

DISCUSSION

This study analyzes sequence data from a North American population of *D. simulans* and documents that the high observed levels of linkage disequilibrium and the chromosomal differences in levels of linkage disequilibrium are not expected under the standard null model. Both demographic and selective departures from the null model are possible explanations, and we considered two of these alternatives to the null model. We describe below some of the difficulties associated with assessing whether these models are appropriate.

The bottleneck model: Not much is known about the demographic history of North American populations of *D. simulans*, but as a human commensal, *D. simulans* is unlikely to have arrived in North America before humans did. The first people in the Americas are thought to have crossed via the Bering Strait ∼14,000–15,000 years ago (see, *e.g.*, Jones *et al.* 1994). If we assume an average of 10 generations a year, then $T_0 = 2000$ generations ago and 1.2×10^5 generations ago correspond to 200 years ago and 12,000 years ago, respectively. However, it seems improbable that *D. simulans* crossed via the Bering Strait, since this would require travel through thousands of miles of harsh Arctic weather and dependence on humans with a low population density. Thus, it may be more likely that *D. simulans* was introduced into the Americas after the European conquest ∼500 years ago. No one knows when *D. simulans* first started crossing the Atlantic as stowaways on ships, but it seems plausible that at first the number of migrants was limited. Both the volume of traffic and the cargo composition changed slowly over time; at some point in the past, successful migration to the Americas must have been possible but difficult. So, independent of genetic data, a recent bottleneck in the history of American populations of *D. simulans* seems to be a reasonable demographic model. We chose to model a single founder event, followed by rapid population growth. Perhaps a more realistic model would have many founder events, spread out over time (continuing to the present day). However, the earliest migrants might have contributed a disproportionally large amount to the gene pool of the new population; the newly founded population may have had ample opportunity to grow, since 500 years ago there were many settled human communities in the Americas. If so, later migrants would then be less important, since they would contribute proportionally very little to the genetic makeup of the population.

In summary, our simple bottleneck model probably captures some fundamental element of the population history of North American *D. simulans*. Assuming that ancestral populations were close to mutation-drift equilibrium, a simple bottleneck model can, at least qualitatively, account for three essential features of the Californian *D. simulans* data: (1) a genome-wide increase in levels of linkage disequilibrium; (2) more linkage disequilibrium on the X than on the autosomes; and (3) a skew in the frequency spectrum toward more common variants on the X relative to the autosomes.

However, this does not necessarily mean that a bottleneck is a sufficient explanation for the patterns of variation in the data analyzed in this article. The effect that a bottleneck has on levels of diversity, linkage disequilibrium, and the frequency spectrum is quite sensitive to many unknown parameters. Exploratory simulations suggest that decreasing $(1 - \delta_\theta)$ or T_0 (while keeping the other parameters constant) leads to a greater increase in linkage disequilibrium, while decreasing n_r leads to more of an effect on the X relative to the autosomes. Also, if the current effective population size and T_0 are larger, there is little effect on levels of linkage disequilibrium. For example, if the current N is 1×10^9, then T_0 must be quite small (*e.g.*, $T_0 \leq 4 \times 10^3$ generations) for a bottleneck to have an appreciable effect on estimates of linkage disequilibrium (results not shown).

Perhaps more worrisome is the fact that the ratio of effective sizes for the X and the autosomes in the ancestral population (n_r) must be low (*i.e.*, ≤ 0.75) to be consistent with the observed ratio of diversities in the Californian population (*i.e.*, ≤ 0.69, the approximate upper bound for $\hat{N}_{\theta x}/\hat{N}_{\theta a}$). In the bottleneck simulations we present (Table 2, Figure 5), we assume $0.6 \leq n_r \leq 0.7$. In other words, we assume that the male effective population size is greater than or equal to the female effective population size. This situation may be unlikely for Drosophila where sexual selection is expected to reduce the effective population size of males relative to females (Crow and Morton 1954). On the other hand, Charlesworth (2001) pointed out that if females are generally in poor breeding condition (as observed by Boulètreau 1987 in an European population), then n_r will be reduced. This may counter the effects of sexual selection in males. In fact, for non-African *D. melanogaster*, Charlesworth (2001) estimates $n_r = 0.73$ and 0.64, with or without sexual selection on males, respectively. Thus, the prebottleneck n_r may have been low if the founding population was non-African (*e.g.*, European). Unfortunately, we have little information about the relative variance in male and female reproductive successes in Drosophila populations or the origin of this particular North American population of *D. simulans*. Under the simplest assumption of equal numbers of males and females (*i.e.*, $n_r = 0.75$), a severe and recent bottleneck can still produce an X/autosome ratio of diversities that is consistent with the findings of Begun and Whitley (2000); if $T_0 = 2000$ generations and $\delta_\theta = 0.75$, then $\hat{N}_{\theta x}/\hat{N}_{\theta a} = 0.684$.

Positive selection models: An alternative to a purely demographic explanation is that natural selection for adaptation has influenced the observed patterns of variation. *D. simulans* originated in Africa (David and Capy

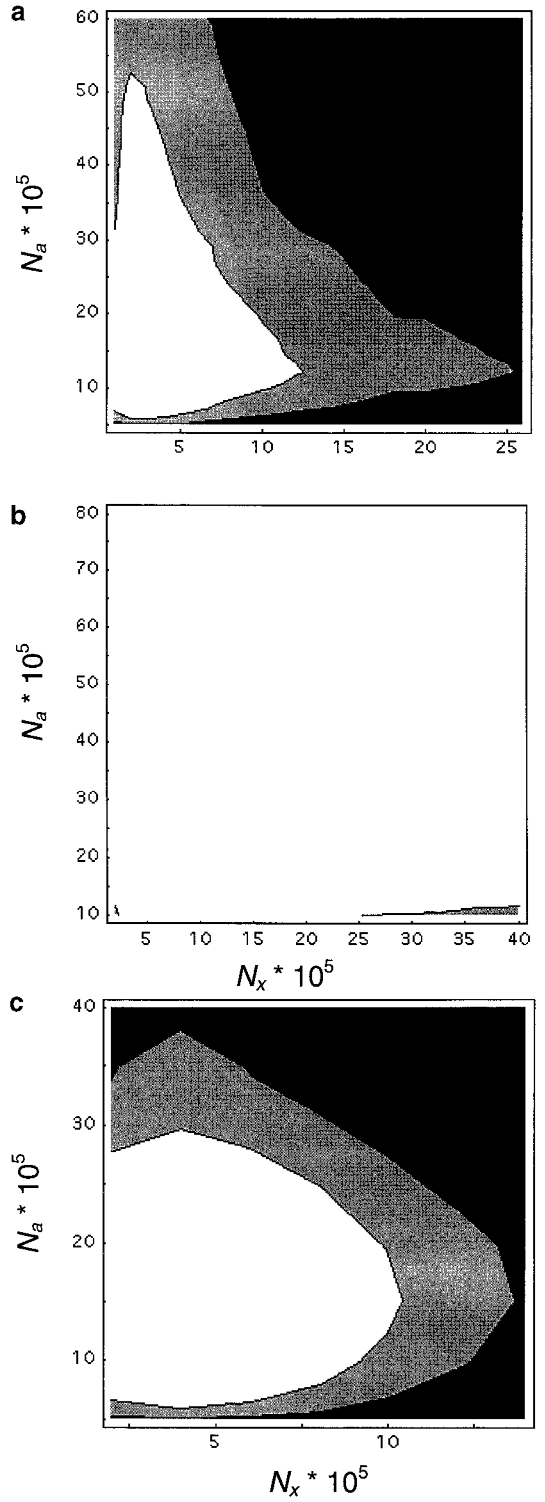

Figure 5.—R^* as a function of N_x and N_a under a model of a recent population bottleneck. (a) $T_0 = 2000$ generations ago, $\delta_\theta = 0.85$, and $n_r = 0.6$. (b) $T_0 = 2000$ generations ago, $\delta_\theta = 0.75$, and $n_r = 0.7$. (c) $T_0 = 1.2 \times 10^5$ generations ago, $\delta_\theta = 0.85$, and $n_r = 0.7$. The shading categories are the same as in Figure 3b.

1988; Lachaise *et al.* 1988), and some adaptive evolution must have occurred while populations coped with different environments and colder climates. We modeled natural selection by simulating recurrent, nonoverlapping selective sweeps linked to a neutral locus. Although under certain assumptions (discussed in Begun and Whitley 2000) this model can reduce X-linked (relative to autosomal) levels of diversity, our simulations show that it is inconsistent with other facets of the data: A simple selective sweep model leads to an excess of rare variants and no appreciable increase in levels of linkage disequilibrium. In contrast, the data show no skew in the frequency spectrum, high levels of linkage disequilibrium on 3R, and extremely high levels of linkage disequilibrium on the X.

We chose the simple recurrent sweep model partly because it has been carefully studied before (*e.g.*, Kaplan *et al.* 1989; Braverman *et al.* 1995) and partly because it is reasonably easy to implement in the coalescent context (where many replicates can be run quickly). However, it is not clear whether the model is appropriate for Drosophila data. For example, our simulations assume that selection is additive, even though one of the main arguments for greater hitchhiking effects on the X invokes dominance effects (see, *e.g.*, Begun and Whitley 2000). This facet is not a major concern, since any process with recurrent, rapid fixation of new alleles is likely to produce a similar pattern in sequence data (*i.e.*, a skew in the frequency spectrum toward rare alleles and no increase in levels of linkage disequilibrium, using the methods in this article).

Another concern is the frequency of selective sweeps. We have chosen simulation parameters that allow few overlapping sweeps. We calculate [similar to (6) in Braverman *et al.* 1995] that the probability that a second selective sweep starts before a given one has finished is >0.05 for values of $N_x \leq 2.4 \times 10^5$ and $N_a \leq 1.5 \times 10^5$. Most of these overlaps consist either of new beneficial alleles arising after an older beneficial allele has already swept to high frequency (but not fixed) and/or two beneficial alleles that are not tightly linked to each other; in both cases, the two sweeps are essentially independent. In general, if s and δ_θ are fixed, then multiple sweeps are more likely to overlap as N decreases. This happens because a selective sweep with a given value of s has an effect on standing levels of linked neutral diversity that is only weakly dependent on N, while sweeps take longer (in units of scaled time) in smaller populations, so are more likely to overlap. Note that we fixed δ_θ so that the effect of selection would be comparable across different values of N. If instead we were to fix the rate of introduction of advantageous alleles, then there would be more sweeps as N increases, and δ_θ would decrease with increasing N; because we have no prior knowledge regarding Λ_r, this implementation does not seem to be appropriate. Since our goal is to determine whether recurrent selective sweeps can produce the excess of linkage disequilibrium that is observed (given the proposed reduction in X-linked *vs.* autosomal diversity), the relevant question is whether larger values of N_x and N_a are compatible with the data. The answer to this question is still no; Figure 4 shows that R^* is very small when both N_x and N_a are large (*i.e.*, when the nonoverlapping sweep assumption is met).

The problem of overlapping sweeps might be exacerbated if the rate of selective events over time is not constant or the strength of selection is weaker. The general effects of a recurrent selective sweep model on the frequency spectrum and levels of linkage disequilibrium are not very sensitive to s, as long as $s \geq 0.002$ (results not shown). However, for smaller selection coefficients (*e.g.*, $s < 0.002$) and the small population sizes considered here, the simple selective sweep model becomes inappropriate due to the large number of overlapping selective events. Also, if natural selection is being driven by adaptation to new environments, then the rate of introduction of favorable alleles might depend heavily on the location and movement of populations and would be much higher at some times than at others. Without any independent source of information on the relevant parameters, we have no idea how often selective sweeps may have overlapped and interfered with each other over time. We also have no idea how multiple competing sweeps (perhaps in a subdivided population) affect levels of variation, the frequency spectrum, or patterns of linkage disequilibrium, or for that matter how sweeps in a subdivided population behave. For any of these models to be viable explanations of the data, they would need to increase levels of linkage disequilibrium on both chromosomes (though much more on the X than the autosomes). They would also need to be able to cause a decrease in levels of variability (on the X) without causing a skew in the frequency spectrum toward rare variants. This seems unlikely unless many of the sweeps are ongoing. Further work will explore how such models affect patterns of sequence polymorphism.

Another possibility is that adaptive evolution operated on standing variation, instead of newly arising mutations. If so, the rate of adaptation on the X might actually be *slower* than the rate on the autosomes (Orr and Betancourt 2001). Nothing is known yet about the predictions of such a model regarding levels of diversity, the frequency spectrum, or levels of linkage disequilibrium on different chromosomes. But, as before, it is unlikely that this model could decrease levels of diversity without affecting the frequency spectrum, unless many selective events have not yet led to fixation of the favored type.

Finally, natural selection might operate in a way that is fundamentally different from the simple directional selection models discussed above. However, Gillespie (1997) examined a range of selective models and found that all had similar effects on levels of diversity and the frequency spectrum (see, *e.g.*, his Figure 3). This makes it less likely that any of them are consistent with the observed frequency spectra and levels of diversity (leaving aside the issue of whether they are consistent with the observed levels of linkage disequilibrium). An alternative put forward by Begun and Whitley (2000) is that the rapid changes in frequency (without fixation) of X-linked meiotic drive or sexually antagonistic alleles may also account for reduced levels of variability on the X relative to the autosomes. Nothing is known about how such a model would affect the frequency spectrum or levels of linkage disequilibrium.

Conclusions: Any evolutionary model that seeks to be a sufficient explanation for the North American *D. simulans* data must simultaneously be consistent with the observed levels of diversity, frequency spectra, and levels of linkage disequilibrium on the X and autosomes. A simple bottleneck model can do so, but only if $n_r \leq 0.75$ and the population size reduction was severe and recent. It is not clear how reasonable these conditions are. On the other hand, a simple hitchhiking model can be rejected because it is inconsistent with both the observed frequency spectra and levels of linkage disequilibrium.

The relative role of natural selection in shaping patterns of *D. simulans* genetic variation remains unknown. More work needs to be done to explore how other models of natural selection affect patterns of variability. These models might examine, *e.g.*, adaptation in structured populations, natural selection in variable environments (*cf.*, Gillespie 1991, 1997; Orr and Betancourt 2001), and/or interference between multiple favorable alleles. This work will give us a sense of which models are plausible for *D. simulans* data.

It will be much easier to test *D. simulans* evolutionary models once sequence polymorphism data from other (predominantly African) populations are gathered. These data might allow one to infer whether migration to the Americas occurred primarily from Europe or from Africa and would help us construct a reasonable demographic null model. Only by explicitly considering demography will we be able to start deciphering the contribution of natural selection for adaptation to different populations of *D. simulans*.

We thank B. Charlesworth and two anonymous reviewers for helpful suggestions on an earlier version of this manuscript. J.D.W. and M.P. were supported by National Science Foundation Postdoctoral Fellowships in Bioinformatics. P.A. was supported by a European Molecular Biology Organization Postdoctoral Fellowship.

LITERATURE CITED

Aguadé, M., N. Miyashita and C. H. Langley, 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657–665.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in Drosophila, pp. 46–56 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

Begun, D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97:** 5960–5965.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail and A. G. Clark, 2000 Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156:** 1879–1888.

Boulètreau, J., 1987 Ovarian activity and reproductive potential in a natural population of *Drosophila melanogaster*. Oecologia **35:** 319–342.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Caballero, A., 1995 On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics **139:** 1007–1011.

Charlesworth, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet. Res. **68:** 131–149.

Charlesworth, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. **77:** 153–166.

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

Crow, J., and N. E. Morton, 1954 Measurement of gene frequency drift in small populations. Evolution **9:** 202–214.

Crow, J. F., and M. J. Simmons, 1983 The mutational load in Drosophila, pp. 1–35 in *The Genetics and Biology of Drosophila*, Vol. 3C, edited by M. Ashburner, H. O. Carson and J. N. Thompson. Academic Press, London.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

Fay, J. C., and C.-I Wu, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution* (Oxford Series in Ecology and Evolution). Oxford University Press, Oxford, UK.

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291–299.

Hamblin, M. T., and M. Veuille, 1999 Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics **153:** 305–317.

Heino, T. I., A. O. Saura and V. Sorsa, 1994 Maps of the salivary gland chromosomes of *Drosophila melanogaster*. Dros. Inf. Serv. **73:** 621–738.

Houle, D., K. A. Hughes, S. Assimacopoulos and B. Charlesworth, 1997 The effects of spontaneous mutation on quantitative traits. II. Dominance of mutations with effects on life-history traits. Genet. Res. **70:** 27–34.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Japan Scientific Society, Tokyo.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

Jones, S., R. Martin and D. Pilbeam (Editors), 1994 *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge, UK.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Lachaise, D., L. M. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

Lemeunier, F., and M. A. Ashburner, 1976 Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond. Ser. B Biol. Sci. **193:** 275–294.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Li, W. H., 1997 *Molecular Evolution*. Sinauer Press, Sunderland, MA.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

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.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Ohnishi, S., and R. A. Voelker, 1979 Comparative studies of allozyme loci in *Drosophila simulans* and *D. melanogaster*. II. Gene arrangement on the third chromosome. Jpn. J. Genet. **54:** 203–209.

Orr, H. A., and A. J. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875–884.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **18:** 291–298.

Sharp, P. M., and W. H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Sniegowski, P., and B. Charlesworth, 1994 Transposable element numbers in cosmopolitan inversions from a natural population in *Drosophila melanogaster*. Genetics **137:** 815–827.

Stephan, W., T. H. E. Wiehe, and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Sturtevant, A. H., 1929 The genetics of *Drosophila simulans*. Carnegie Inst. Wash. **399:** 1–62.

Tajima, F., 1989a Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tajima, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

Takano-Shimizu, T., 1999 Local recombination and mutation effects on molecular evolution in Drosophila. Genetics. **153:** 1285–1296.

True, J. R., J. M. Mercer and C. C. Laurie, 1996 Differences in frequency and distribution among three sibling species of Drosophila. Genetics **142:** 507–523.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wall, J. D., and R. R. Hudson, 2001 Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. **18:** 1134–1135.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: H. Ochman