Demography and Natural Selection Have Shaped Genetic Variation in *Drosophila melanogaster*: A Multi-locus Approach

Sascha Glinka,[1] Lino Ometto,[1] Sylvain Mousset, Wolfgang Stephan and David De Lorenzo[2]

Section of Evolutionary Biology, Department of Biology II, University of Munich, D-80333 Munich, Germany

Manuscript received June 24, 2003
Accepted for publication July 23, 2003

ABSTRACT

Demography and selection have been recognized for their important roles in shaping patterns of nucleotide variability. To investigate the relative effects of these forces in the genome of *Drosophila melanogaster*, we used a multi-locus scan (105 fragments) of X-linked DNA sequence variation in a putatively ancestral African and a derived European population. Surprisingly, we found evidence for a recent size expansion in the African population, *i.e.*, a significant excess of singletons at a chromosome-wide level. In the European population, such an excess was not detected. In contrast to the African population, we found evidence for positive natural selection in the European sample: (i) a large number of loci with low levels of variation and (ii) a significant excess of derived variants at the low-variation loci that are fixed in the European sample but rare in the African population. These results are consistent with the hypothesis that the European population has experienced frequent selective sweeps in the recent past during its adaptation to new habitats. Our study shows the advantages of a genomic approach (over a locus-specific analysis) in disentangling demographic and selective forces.

IN the past decade, evidence that natural selection plays a key role in shaping genome-wide patterns of variability in Drosophila has been mounting (Aquadro 1997). However, it remains a challenge to discern selection from other forces, particularly demographic factors. Only recently, studies have begun to address this problem by consistently sampling populations and using multiple loci (Begun and Whitley 2000). The rationale of this approach is that demographic processes affect the entire genome in a similar way, whereas selective forces tend to leave locus-specific footprints that are detectable in a genome-wide survey.

Drosophila melanogaster, originating from sub-Saharan Africa, is believed to have expanded its range after the last glaciation (*i.e.*, ~10,000–15,000 years ago; David and Capy 1988; Lachaise *et al.* 1988). During this habitat expansion, demographic processes (such as bottlenecks and subsequent population size increases) would be expected to have occurred. In addition, selective events are likely to have played an important role in the adaptation of this species to its new environments.

To distinguish demographic and selective processes important for the recent adaptations of *D. melanogaster*, we compared a putatively ancestral population from Africa (Zimbabwe) with a derived population from Europe (The Netherlands). Since a whole-genome scan of DNA sequence variation is currently not feasible, we used a multi-locus approach. The availability of the genomic sequence of *D. melanogaster* made this approach possible. To be able to discern different selective regimes, we focused on chromosomal regions of normal recombination (Kim and Stephan 2002). Furthermore, we used sequence variation rather than microsatellites (Harr *et al.* 2002) for the following reasons. One of our long-term goals is to estimate the rate of advantageous substitutions in the recent past of *D. melanogaster*. Advantageous substitutions causing sweeps that have occurred no more than ~0.1 N_e (effective population size) generations ago can be detected with sufficiently high power using single nucleotide polymorphisms (Kim and Stephan 2000; Przeworski 2002). For *D. melanogaster*, 0.1 N_e generations correspond to ~10,000–15,000 years. This window of time matches very well the colonization of Europe by *D. melanogaster*. Thus, the use of DNA sequence variation should enable us to detect most of the sweeps that have occurred during this colonization period and hence to obtain a reliable estimate of the rate of advantageous substitutions. In contrast, with microsatellites that mutate faster than nucleotides we may be able to observe only the very recent sweeps. Since this is the first screen of DNA sequence variation in *D. melanogaster*, we concentrated on the X chromosome.

MATERIALS AND METHODS

Population samples: *D. melanogaster* data were collected from 24 highly inbred lines derived from two populations: 12 lines from Africa (Lake Kariba, Zimbabwe; Begun and Aquadro 1993) and 12 lines from a European population (Leiden, The Netherlands). The Zimbabwean lines were

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Section of Evolutionary Biology, Department of Biology II, University of Munich, Luisenstrasse 14, D-80333 Munich, Germany. E-mail: delorenzo@lmu.de

kindly provided by C. F. Aquadro, and the European ones were provided by A. J. Davis. Furthermore, a single *D. simulans* inbred strain (Davis, CA; kindly provided by H. A. Orr) was used for interspecific comparisons.

PCR amplification and DNA sequencing: On the basis of the available DNA sequence of the *D. melanogaster* genome (FlyBase 2000, Release 2, http://www.flybase.org), we amplified and sequenced 105 fragments of noncoding DNA (from 63 introns and 42 intergenic regions), randomly distributed across the entire euchromatic portion of the X chromosome. Most fragments are located in regions of intermediate to high recombination rates. However, 11 fragments are from the telomeric region exhibiting low recombination rates, *i.e.*, distal to the *white* locus (see online supplemental Tables 1 and 2 available at http://www.genetics.org/supplemental). We amplified and sequenced the homologous 105 fragments in a single strain of *D. simulans*.

We extracted genomic DNA from 10 females of each inbred line using the PUREGENE DNA isolation kit (Gentra Systems, Minneapolis). The PCR products were then purified with EXOSAP-IT (USB, Cleveland). Sequencing reactions were performed for both strands according to the protocol of the DYEnamic ET terminator cycle sequencing kit (Amersham Biosciences, Buckinghamshire, UK) and run on a MegaBACE 1000 automated capillary sequencer (Amersham Biosciences). Analysis of the data was done using the software Cimarron 3.12 (Amersham Biosciences) for lane tracking and base calling. Only good-quality sequences (MegaBACE quality score of at least 95 out of 100) were aligned and checked manually with the application Seqman of the DNAstar (Madison, WI) package. Singletons were confirmed by reamplification and resequencing. The sequences were deposited in the EMBL database (for accession numbers, see online supplemental information at http://www.genetics.org/supplemental).

Statistical analysis: Basic population genetic parameters were estimated with the program DnaSP 3.98 (Rozas and Rozas 1999). Levels of nucleotide diversity were estimated using π (Tajima 1983) and θ (Watterson 1975). For this analysis, we considered the total number of mutations rather than the number of segregating sites, because in a few instances we observed three different nucleotides segregating at the same position.

To test the neutral equilibrium model, we employed the multi-locus Hudson-Kreitman-Aguadé (HKA) and Tajima's *D* tests (Hudson *et al.* 1987; Tajima 1989). Both tests were done using the program HKA, kindly provided by J. Hey (http://lifesci.rutgers.edu/heylab), in which the test statistics were compared with the distributions generated from 10,000 coalescent simulations (Kliman *et al.* 2000).

In addition, we used the following statistics: the number of haplotypes, *K*, and the haplotype diversity, *H* (Depaulis and Veuille 1998), and, for the African population, Fay and Wu's *H* (Fay and Wu 2000). These statistics were calculated with the program DnaSP 3.98 (Rozas and Rozas 1999). We generated the empirical distributions of these statistics for each fragment using coalescent simulations (10,000 iterations; Hudson 1990, 1993), conditioned on the number of segregating sites (Depaulis *et al.* 2001), and a population recombination rate, *R* (programs are available from S. Mousset). Since in *D. melanogaster* there is no recombination in males, the population recombination rate, *R*, was estimated by $2N_e c$, where c is the female recombination rate per fragment per generation (Przeworski *et al.* 2001). N_e was assumed to be 10^6 (Li *et al.* 1999), and, for each fragment, c was estimated by multiplying the per-site-recombination rate, r (see below), by its length, *L*.

Recombination rate: We estimated r (recombination rate per site per generation) for each fragment as follows. We used a computer program of Comeron *et al.* (1999) to obtain an estimate of the recombination rate for each fragment. This algorithm follows the method of Kliman and Hey (1993). We compared our results to two other estimators of the recombination rate: the adjusted coefficient of exchange (ACE; Begun and Aquadro 1992) and the procedure proposed by Charlesworth (1996).

For the latter method, we used the absolute position of each fragment to calculate physical distances. The estimate of the recombination rate is therefore expressed in centimorgans per megabase instead of centimorgans per band (see Charlesworth 1996). We divided the X chromosome into two regions containing all of our 105 fragments: (I) the distal-*white* region (0.2–2.45 Mb, 0.02–1.5 cM) and (II) the proximal-*white* region (2.45–16.89 Mb, 1.5–56.7 cM). Following Charlesworth (1996), the *white* locus (2.45 Mb, 1.5 cM) was chosen as a transition point between region I and region II.

Demographic modeling of the European population: Because extant European *D. melanogaster* are believed to be derived from an ancestral African population (David and Capy 1988), we tested the observed data against simple demographic null models: (i) a constant-population-size model and (ii) a population-size-bottleneck model with subsequent expansion (Wall *et al.* 2002; Lazzaro and Clark 2003). In the latter model, we simulated a population of initial effective size N_i, crashing T_b generations ago to size N_b. After T_m generations, the population was allowed to grow exponentially to the current effective population size, N_0.

The following parameters had to be specified for each fragment: the mutational parameter, θ (estimated from data); the sample size, n; and the fragment length, *L*. Constant-population-size models were tested using the observed average θ value of the European population, while the bottleneck models were conditioned on the observed average θ value of the African population (*i.e.*, the value of the hypothetical ancestral population). Our simple models assumed no intragenic recombination but did assume free recombination between fragments. We used several combinations of values of N_b, N_0/N_i, and T_b. T_m was adjusted to obtain a total number of segregating sites in a simulation close to the observed value of 737. For each fragment, 10,000 genealogies were simulated using the program "ms" (Hudson 2002) under the demographic models mentioned above. The probability of observing exactly $F_0 = 13$ fragments with no polymorphism in our simulation (see RESULTS) was then calculated as the proportion of simulated samples with exactly 13 fragments with no polymorphic sites. This probability was used in a two-tailed likelihood-ratio test as a likelihood of our observation; when the probability was $<10^{-4}$, we used 10^{-4} as a conservative overestimate of this value.

RESULTS

DNA sequences for 105 X chromosome fragments were obtained from 10–12 lines of an African and a European population of *D. melanogaster* (with an average of 11.9 lines per sample). The size of the fragments varied between 240 and 781 bp (excluding insertions and deletions) with a mean (SE) of 517 bp (11 bp). The total region from which these fragments derive spans ~14 Mb. This results in an average distance between adjacent fragments of ~140 kb (Figure 1).

There are several large gaps in our genome scan (Figure 1), in which we could not recover a sufficient num-

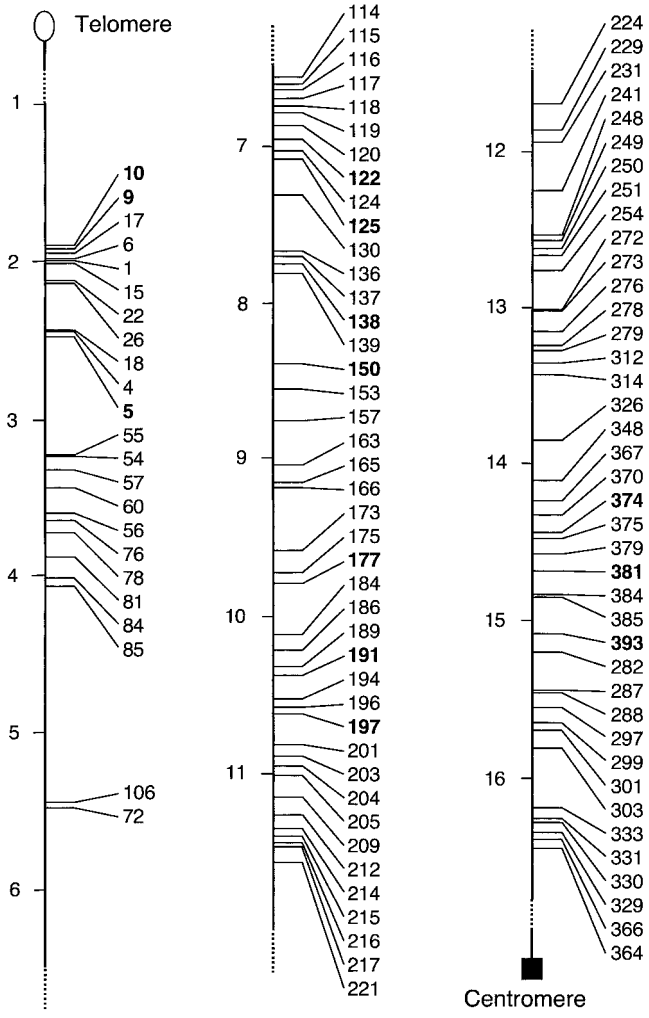

FIGURE 1.—Distribution of the sequenced fragments along the X chromosome. Fragments are shown by their absolute position (distances in megabases from the telomere). Fragments with no polymorphism in the European sample are in boldface type.

ber of sequences (*i.e.*, at least 10 per sample and the sequence of the *D. simulans* line). The majority of fragments (103) are located in two segments (between coordinates 1.9 and 4.1 Mb and between 6.5 and 16.4 Mb from the telomere, respectively), thus spanning a region of 12 Mb with an average distance of 119 kb between fragments. The region between these two segments appears to contain a high density of repetitive DNA (for instance, microsatellites; HARR *et al.* 2002) that may have caused problems with PCR and sequencing. The details are being investigated.

In both *D. melanogaster* samples, intergenic regions and introns did not produce significantly different results when analyzed separately (results not shown) and are therefore pooled in the following analyses.

Polymorphism patterns in the African population: Figure 2, a–c, and Table 1 in the online supplemental information at http://www.genetics.org/supplemental provide a summary of the polymorphism and divergence data. Of the 54,944 sites sequenced (excluding insertions and deletions), 2057 are polymorphic. The mean of θ (SE) is 0.0127 (0.0007), which is higher than the average value of 0.0071 reported for noncoding regions on the *D. melanogaster* X chromosome (MORIYAMA and POWELL 1996), but lower than the average value of 0.0257 estimated for synonymous X-linked sites for African populations from diverse geographic localities (ANDOLFATTO 2001). For π, the result is similar: 0.0112 (0.0007) to 0.0074 (MORIYAMA and POWELL 1996) and 0.0242 (ANDOLFATTO 2001).

We tested our data for compatibility with the neutral equilibrium model. The HKA test is used to determine whether the levels of intraspecific polymorphism and interspecific divergence at our set of fragments are consistent with the equilibrium model (HUDSON *et al.* 1987). A multi-locus version of the original HKA test was applied to all 105 fragments in the African sample (Figure 3a). No significant departure from the equilibrium model was detected ($\chi^2 = 93.31$, $P = 0.765$).

We also calculated the Tajima's *D* statistic for each fragment and tested whether the observed average across fragments was consistent with the equilibrium model by estimating the critical values of this distribution from coalescent simulations (see MATERIALS AND METHODS). In these simulations, we assumed no intragenic recombination (but free recombination between fragments). The African population shows a negative average value (SE) of Tajima's *D* of −0.578 (0.058). None of the 10,000 simulated samples of 105 fragments had a more extreme average value of *D*. This suggests that our data depart from the neutral equilibrium model. In fact, most of the fragments have negative *D* values (sign test, two-tailed, $P < 0.001$; Figure 2d).

To further investigate the pattern of variation in the African sample, we focused on two statistics, the number of haplotypes, *K*, and the haplotype diversity, *H* (DEPAULIS and VEUILLE 1998). Low values of these statistics indicate that there are too few haplotypes in the sample due to demographic (*e.g.*, population substructure and/or weak bottlenecks) and/or selective events (*e.g.*, incomplete hitchhiking; DEPAULIS and VEUILLE 1998). On the other hand, high values can result from population expansion or old, complete hitchhiking events (DEPAULIS and VEUILLE 1998). Because recombination tends to increase both statistics, we used the estimated recombination rate (COMERON *et al.* 1999; see MATERIALS AND METHODS) for each fragment in the coalescent simulations. Assuming that this recombination rate is correct, we can perform a two-tailed test. Under neutrality, we expect an equal proportion of the observed values to be lower and higher than the simulated median.

We found that the observed haplotype diversity, *H*, was higher than the simulated median in 78 of the 105 fragments; this proportion is significantly larger than

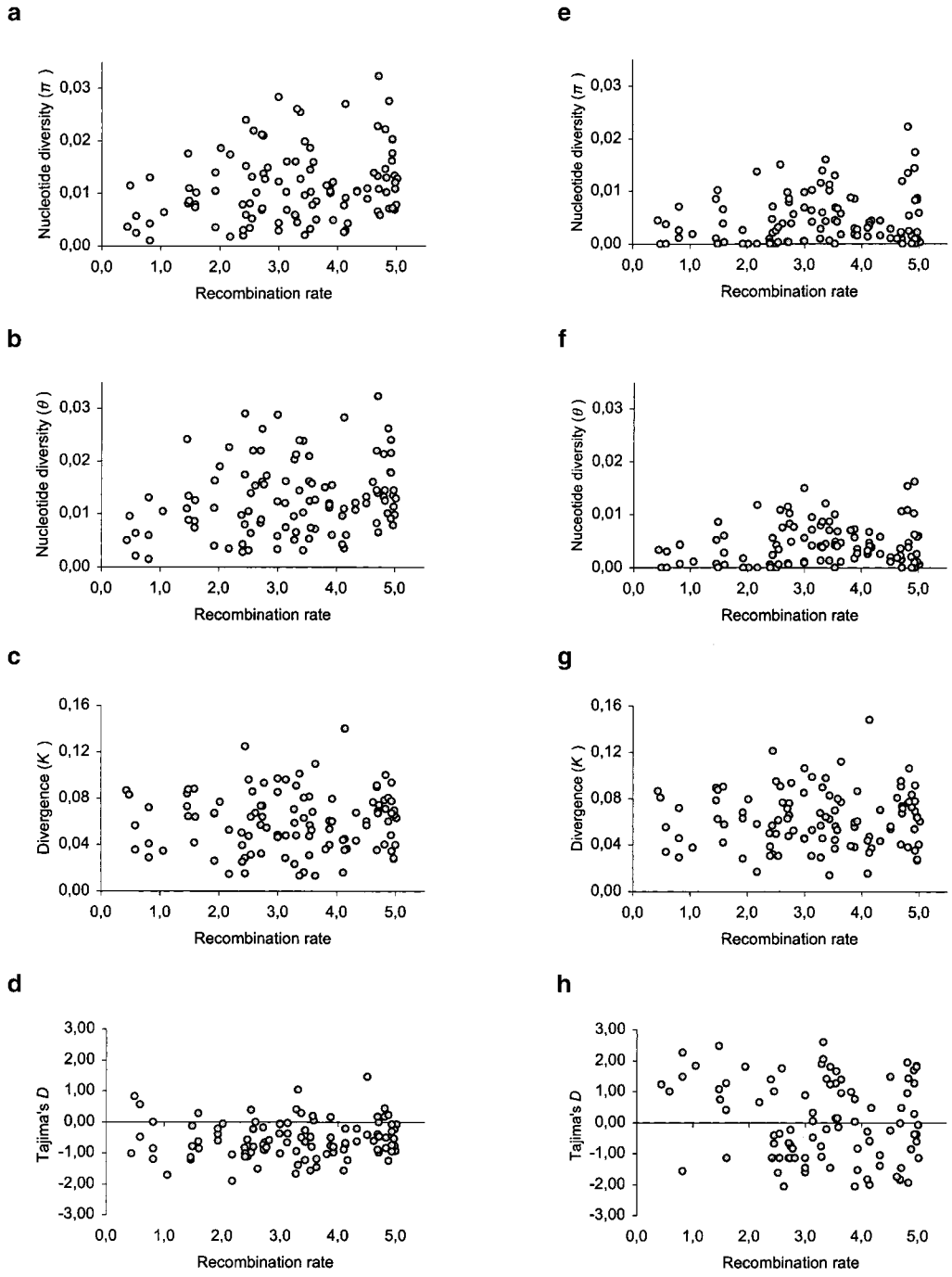

FIGURE 2.—Nucleotide diversity π and θ, divergence *K*, and Tajima's *D* *vs.* recombination rate. (a–d) African population. (e–h) European population. Recombination rate is expressed in recombination events per site per generation $\times 10^8$ (COMERON *et al.* 1999).

expected (sign test, two-tailed, $P < 0.001$). For the number of haplotypes, *K*, a significant trend toward a higher number was also observed (sign test, two-tailed, $P = 0.03$). High values of haplotype diversity and large numbers of haplotypes can result from a star-like genealogy due to population expansion or complete hitchhiking events (DEPAULIS and VEUILLE 1998).

Assuming that recurrent complete selective sweeps occur along a recombining chromosome, we expected to detect the footprints of partial sweeps as well. We thus examined whether there is evidence for partial hitchhiking events using the *K*- and *H*-haplotype tests (DEPAULIS and VEUILLE 1998) and Fay and Wu's *H* test (FAY and WU 2000). Since we were exploring possible departures of these statistics at their lower bounds, we used the conservative assumption of zero recombination (DEPAULIS and VEUILLE 1998). For the 105 fragments, we observed only one significant Fay and Wu's *H* value (one-tailed, $P = 0.03$).

These results, together with the observations from the HKA test, argue against a model of recurrent selective sweeps (BRAVERMAN *et al.* 1995) as an explanation of the chromosome-wide excess of singletons observed in the African population. It appears that this pattern of

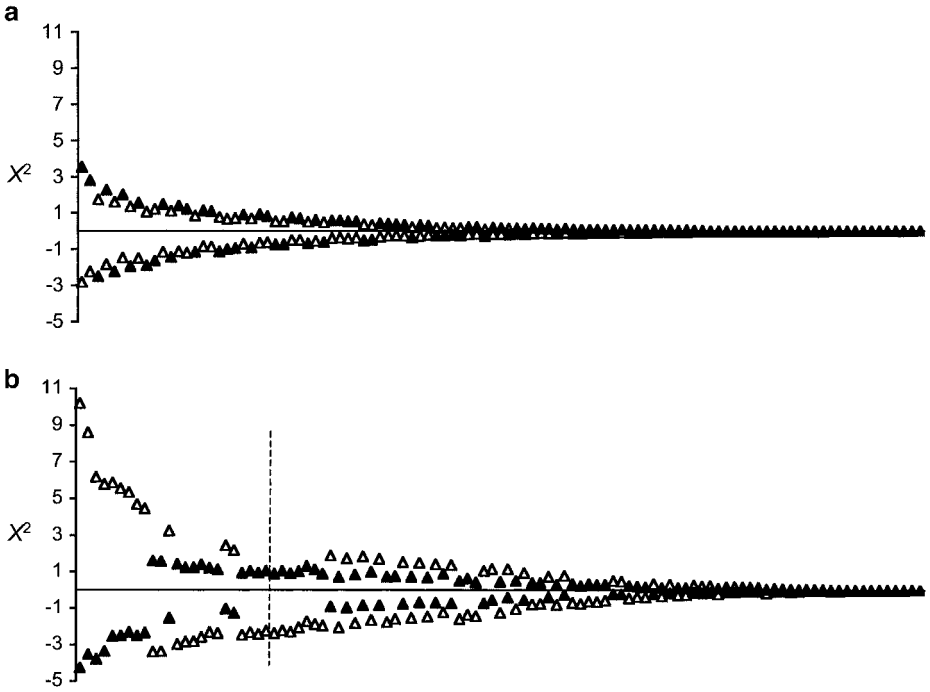

FIGURE 3.—Contribution of each fragment to multi-locus HKA statistic. (a) African population and (b) European population. For each fragment, the contributions to the overall test statistic by the polymorphism (△) and divergence (▲) data are shown. Values above (below) the *x*-axis indicate a larger (smaller) contribution than expected. Fragments are ranked along the *x*-axis according to their total contribution to the test statistic (including polymorphism and divergence components). When the 24 fragments at the left of the vertical dashed line were excluded from the test (for the European sample), the value of the overall test statistic dropped below the critical value.

polymorphism has most likely been shaped by demography.

Is there any evidence for a signature of selection in the African population? Using two-tailed tests, we found a (weak) positive correlation between recombination rate and nucleotide variation (as measured by π and θ; see Figure 2, a and b): for π, Pearson's $R = 0.246$, $P < 0.02$, Spearman's $R = 0.237$, $P < 0.02$; for θ, Pearson's $R = 0.237$, $P < 0.02$, Spearman's $R = 0.234$, $P < 0.02$. If this observation were due to a lower neutral mutation rate in regions of reduced recombination, then these regions should also be less diverged. However, we found no correlation between recombination rate and levels of divergence (Pearson's $R = 0.003$, $P > 0.10$, Spearman's $R = 0.028$, $P > 0.10$; Figure 2c). If we consider only fragments above a certain recombination rate (for example, 2×10^{-8} recombination events per base pair per generation, which corresponds to our previously defined region II; see MATERIALS AND METHODS), thus including 94 loci, then the correlation between recombination rate and polymorphism disappears (for π, Pearson's $R = 0.158$, $P > 0.10$; for θ, Pearson's $R = 0.115$, $P > 0.20$). These conclusions hold for all three measures of recombination rates (see MATERIALS AND METHODS), except that the (weak) correlation between nucleotide diversity and ACE was still found when the 11 fragments located in regions of low recombination were excluded (Pearson's $R = 0.203$, $P < 0.05$, and Pearson's $R = 0.199$, $P < 0.05$ for π and θ, respectively). This suggests that the strong positive correlation between recombination rates and nucleotide diversity reported in previous studies is attributable mainly to loci in low recombination regions (BEGUN and AQUADRO 1992; AQUADRO *et al.* 1994; ANDOLFATTO and PRZEWORSKI 2001).

Polymorphism patterns in the European population: A summary of the polymorphism and divergence data is shown in Figure 2, e–g. Of the 55,150 sites sequenced, 737 are polymorphic. The number of segregating sites and estimates of nucleotide diversity for each fragment are shown in the online supplemental Table 2 available at http://www.genetics.org/supplemental/. The means (SE) of π and θ across the X chromosome are 0.0046 (0.0005) and 0.0044 (0.0004), respectively.

In Figure 2, e and f, the estimates of π and θ are plotted against the recombination rate. We observed no significant correlation between nucleotide diversity and any of the three estimates of the recombination rate (MATERIALS AND METHODS). With regard to the first of these recombination rate estimates, the results of the correlation analysis are as follows (two-tailed tests): Pearson's $R = 0.150$ and 0.180 with $P > 0.12$ and $P > 0.06$ for π and θ, respectively; Spearman's $R = 0.137$ and 0.183 with $P > 0.16$ and $P > 0.06$. Also, no correlation between recombination rate and divergence was observed (Figure 2g; Pearson's $R = 0.035$, $P > 0.73$, Spearman's $R = 0.021$, $P > 0.82$). These results contradict to some extent our findings in the African sample, where a weak positive correlation between recombination rate and levels of variation was detected. Since this correlation has been proposed to be an effect of selection (MAYNARD SMITH and HAIGH 1974; CHARLESWORTH 1996), it may indicate that selection in the European population is not as strong as in the African population, perhaps due to interfering demographic processes.

TAJIMA's (1989) test was applied to the European

sample as described in MATERIALS AND METHODS. The observed average of Tajima's *D* (SE) across fragments is 0.045 (0.574). The average value is not significantly different from zero, but the standard error is ($P < 0.0001$). Does this mean that the European population is in equilibrium with regard to demographic and selective forces? Several lines of evidence speak against this hypothesis. Although the mean of Tajima's statistic is close to zero, for 11 fragments the data are not compatible with the neutral equilibrium model. The Tajima test (in its single-locus version; TAJIMA 1989) revealed seven fragments with significantly negative *D* values and four with positive ones. Inspection of the data shows that Tajima's *D* is negative in the fragments exhibiting a rare haplotype with many singletons or strongly positive when most of the variants are organized in a few common haplotypes (Figure 2h). As a result of this, it appears that the mean of *D* across fragments does not differ from zero.

Using the same approach as for the African population sample, we computed the distribution of the *H*- and *K*-haplotype statistics (DEPAULIS and VEUILLE 1998) and recorded the proportion of observed values that were lower and higher than the simulated median. The observed *H* values were lower than the simulated median for 83 fragments; this proportion is higher than expected (sign test, two-tailed, $P < 0.0001$). For *K*, the trend toward fewer haplotypes was also significant (sign test, two-tailed, $P < 0.005$). In agreement with this observation, we found 13 fragments with a significantly low value of *K* or *H*, using the conservative assumption of no recombination in one-tailed *K* or *H* tests. These observations are consistent with the occurrence of bottlenecks and/or selective events in the recent past.

To further investigate whether the data deviate from the neutral equilibrium model, we used the multi-locus version of the HKA test (MATERIALS AND METHODS). A significant departure of the data from this model was detected ($\chi^2 = 238.28$, $P = 0.0016$). Figure 3b shows the contributions of each fragment to the summary statistic (see also online Table 3 at http://www.genetics.org/supplemental/ for details). Furthermore, Figure 3b depicts whether the observed polymorphism and divergence values are lower or higher than expected. The HKA test was repeated with the exclusion of just those fragments with the strongest departures from expectation. The value of the overall test statistic dropped below the critical value at which the test was no longer significant, if 24 fragments with the largest contributions were removed (data not shown; 12 of these fragments show an excess of polymorphism, and 12 a deficiency of polymorphism; see Figure 3b). Note that some of these low-polymorphism fragments contribute to the overall test statistic to a very similar degree as the ones following at higher ranks; *i.e.*, between the fragments at rank 20 and at rank 30 the per-fragment contribution differs by <0.5. All these fragments have values of $\theta \leq 0.0011$.

Next we analyze the fragments exhibiting low levels of variation. In our survey, 13 fragments had no polymorphic sites at all (Figure 1 and online Table 2 at http://www.genetics.org/supplemental/). Furthermore, 12 low-variation fragments have been identified by the HKA test, including 8 of the nonpolymorphic fragments and 4 with extremely reduced nucleotide variability ($\theta \leq 0.0007$).

We first concentrate our analysis on the set of fragments with zero polymorphisms. We used coalescent simulations to test the hypothesis that simple demographic null models (see MATERIALS AND METHODS) can explain our observation of 13 fragments with zero polymorphisms. These are a neutral model of constant population size and various bottleneck models (Table 1). Since the European population is believed to be derived from Africa (DAVID and CAPY 1988; ANDOLFATTO 2001), the prebottleneck effective population size (N_i) is assumed to be equal to the effective size of the Zimbabwean population (*i.e.*, $\sim 10^6$). Different values of N_0 for the European population (between 0.25 and 0.5 N_i)—accounting for the fact that the observed θ value in the European population is about one-third of the estimate of the African population—were assumed. Severe bottlenecks were introduced mimicking the founding of the European *D. melanogaster* population. The values of the parameters (describing the time of occurrence, severity, and duration of a bottleneck) were chosen such that the current simulated population has about the same number of segregating sites as observed.

Among the models tested, a likelihood-ratio two-tailed test shows that some models fit the observation of 13 fragments with no polymorphism better than the neutral (constant population size) model [*e.g.*, bottleneck (Bot) 10, $G = 14.1$, $P = 0.014$, see Table 1]. Appreciable probabilities of getting at least 13 fragments with no polymorphic sites were obtained only for parameter values of the bottleneck model in which the effective population size recovered to its current size in a relatively short time period ($\sim 0.1\ N_0$ generations). Other more realistic scenarios, in which the European population was founded 10,000–15,000 years ago, corresponding to $> \sim 100{,}000$ generations (DAVID and CAPY 1988; LACHAISE *et al.* 1988), and grew more slowly to its current effective size, appear to be inconsistent with our observation of 13 fragments with no polymorphism.

Further evidence against a simple model of population founding followed by expansion is provided by the last two columns of Table 1. First, the average value of Tajima's *D* is negative in all simulations of the bottleneck model. Second, very few simulation runs produced values of Tajima's *D* greater than the observed value (across fragments).

Comparison of the African and European populations: The European population shows lower levels of variation than the African population shows (see above). These differences are statistically significant (Wilcoxon

TABLE 1

Demographic modeling of the European population

Model	Model parameters				$\bar{F}_0$	P ($F_0 \leq 13$)	P ($F_0 = 13$)	P ($F_0 \geq 13$)	Average $\bar{D}$	P ($\bar{D} \geq 0.045$)
	T_b	N_b	T_m	N_0/N_i						
Constant	—	—	—	—	1.26	1	$<10^{-4}$	$<10^{-4}$	−0.077	0.0847
Bot 1	100,000	1,000	3,600	0.5	2.60	1	$<10^{-4}$	$<10^{-4}$	−0.967	$<10^{-4}$
Bot 2	100,000	1,000	7,500	0.25	0.60	1	$<10^{-4}$	$<10^{-4}$	−1.050	$<10^{-4}$
Bot 3	100,000	500	1,750	0.5	2.50	1	$<10^{-4}$	$<10^{-4}$	−0.955	$<10^{-4}$
Bot 4	100,000	500	4,150	0.25	0.55	1	$<10^{-4}$	$<10^{-4}$	−1.049	$<10^{-4}$
Bot 5	50,000	1,000	2,900	0.5	9.14	0.9336	0.0512[a]	0.1176	−0.672	$<10^{-4}$
Bot 6	50,000	1,000	4,400	0.25	3.13	1	$<10^{-4}$	$<10^{-4}$	−1.028	$<10^{-4}$
Bot 7	50,000	500	1,500	0.5	9.08	0.9314	0.0484[a]	0.1167	−0.712	$<10^{-4}$
Bot 8	50,000	500	2,250	0.25	2.94	1	$<10^{-4}$	$<10^{-4}$	−1.049	$<10^{-4}$
Bot 9	25,000	1,000	2,750	0.5	22.40	0.0132	0.0070	0.9938	−0.355	$<10^{-4}$
Bot 10	25,000	1,000	3,850	0.25	12.51	0.6333	0.1153[a]	0.4820	−0.790	$<10^{-4}$
Bot 11	25,000	500	1,300	0.5	20.21	0.0440	0.0210	0.9770	−0.335	0.0013
Bot 12	25,000	500	2,000	0.25	11.56	0.7407	0.1093[a]	0.3696	−0.850	$<10^{-4}$

The models are denoted as follows: Constant, constant population size without recombination; Bot 1–12, bottleneck models without recombination for 12 different sets of values of T_b, N_b, and N_0/N_i. A severe bottleneck of size N_b was introduced T_b generations ago in a population of initial size N_i and maintained for T_m generations. After that time, the population was allowed to grow exponentially to the current population size N_0. $N_i = 10^6$ was assumed. The value of the population mutation parameter was 0.0127, which is equal to the observed average value of θ for the African sample. For the constant-size simulations, the corresponding θ value of the European sample was used. The values of T_m were chosen such that the simulated and observed total numbers of segregating sites across all 105 fragments are in close agreement. F_0 is the number of fragments with no variation; P ($F_0 \leq 13$), P ($F_0 = 13$), and P ($F_0 \geq 13$) are the probabilities of obtaining at most, exactly, or at least 13 fragments with no polymorphism, respectively. Average $\bar{D}$ is the value of Tajima's D across all fragments averaged over all 10,000 simulation runs, and P ($\bar{D} \geq 0.045$) is the probability of observing a value of Tajima's $\bar{D}$ across fragments equal to or larger than the value observed in the European sample.

[a] Likelihood-ratio test, two-tailed, $P < 0.05$ (*i.e.*, the respective bottleneck model fits the observation of $F_0 = 13$ better than the Constant).

matched-pairs signed-ranks test, two-tailed, $P < 0.0001$ for both π and θ). As evident from the larger difference in the means of θ (relative to those of π), the African population harbors more rare variants than the European population does. This is also suggested by the significantly negative average value of Tajima's D for the African population, whereas, in the European population, average D is close to zero.

A large proportion (65%) of the polymorphisms in the European population are also present in the African one (comprising ~23% of the variation found in the African population). This result supports the African origin of the European population. Nonetheless, both populations are considerably differentiated: average F_{ST} (SE; Hudson *et al.* 1992) across fragments is 0.293 (0.017; see online Table 3 at http://www.genetics.org/supplemental/).

Because of suggestions in the literature of differential migration patterns of neutral and selected loci (Cavalli-Sforza 1966; Lewontin and Krakauer 1973), we have investigated differentiation across fragments in more detail. However, instead of using the F_{ST} approach (which was questioned by many authors, *e.g.*, Nei and Maruyama 1975 and Robertson 1975), we asked (more directly) whether derived variants that are fixed in the European sample are in high frequency in the African sample. If this were the case, the colonization history of Europe by African *D. melanogaster* may be explained by a combination of demographic processes and genetic drift without invoking selection.

For each fragment, we recorded the frequency of the derived variants in the African and the European samples. A variant was classified as ancestral when present also in *D. simulans*; when neither of the two *D. melanogaster* variants was found in *D. simulans*, the segregating site was not considered. A total of 1974 segregating sites were classified, including shared polymorphisms, population-specific polymorphisms, and fixed differences. The fragments were partitioned into two groups: (i) those with very low polymorphism [using the HKA test, this was defined in two ways (see above); independent of this definition, however, this group contained the fragments with zero polymorphisms] and (ii) the rest of the fragments. Our results of the HKA test suggest classifying a fragment as a low-variation fragment if (a) $\theta \leq 0.0007$ (21 fragments) or if (b) $\theta \leq 0.0011$ (29 fragments).

Figure 4 compares the relative number of segregating sites for each frequency class for the low-variation fragments defined by criterion b; criterion a gave similar

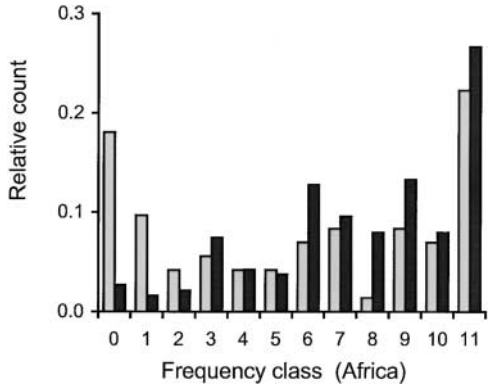

FIGURE 4.—Relative count of derived variants that are fixed in the European sample against their frequency in the African sample. The count of variants across frequency classes is normalized to one. Shaded bars denote variants found in low-variation fragments; solid bars denote variants in the rest of the fragments. Low-variation fragments are defined by criterion b (see *Comparison of the African and European populations*).

results (data not shown). In this analysis, a total of 260 segregating sites with the variant fixed in the European population sample have been used (53 and 72 in the low-variation fragments for a and b, respectively). In the fragments with low variation, there is an excess of derived variants that are fixed in the European sample and rare in the African population. The difference between the low-variation fragments and the rest of the fragments is highly significant. If all 12 frequency classes are considered separately, $\chi^2 = 28.72$, d.f. = 11, $P = 0.0025$, and $\chi^2 = 32.39$, d.f. = 11, $P = 0.0007$ for a and b, respectively; if the low-frequency classes "0" and "1" are lumped together into a single category, leaving all the other classes as the second category, $\chi^2 = 25.19$, d.f. = 1, $P < 0.0001$, and $\chi^2 = 26.42$, d.f. = 1, $P < 0.0001$ for a and b, respectively.

A neutral model, in which the European variants were "sampled" from the African pool and, after colonization, reached high frequency by drift, cannot explain the observed excess of derived variants that are fixed in the low-variation fragments of the European sample and in low frequency in Africa. This observation is consistent with the result that the European population is significantly more diverged from *D. simulans* than is the African population (Wilcoxon matched-pairs signed-ranks test, two-tailed, $P < 0.001$).

DISCUSSION

Our genomic scan of X-linked variation in an African and a European *D. melanogaster* population provides evidence for the impact of demography and natural selection in the recent past during which this species expanded its range. The main features of our data are discussed below.

Demography: Our findings that levels of polymorphism are higher in the African population and that the majority of the sites segregating in the European population are also polymorphic in the African sample confirm previous results (BEGUN and AQUADRO 1993, 1995; ANDOLFATTO 2001). Furthermore, our results are consistent with the hypothesis that *D. melanogaster* originated in sub-Saharan Africa before spreading to the rest of the world (DAVID and CAPY 1988; LACHAISE *et al.* 1988).

A surprising observation, however, was that the African population shows a signature of a recent population size expansion, *i.e.*, a significant excess of singletons at a chromosome-wide level. The reason for this population size expansion remains unclear. Since we found only very little evidence for selective adaptations in the African population (see below), the population size increase does not appear to mirror a change of or an expansion to a new habitat.

The demographic processes that have occurred in the European population are more complex. Our observation that a large number of loci have strongly positive and negative *D* values (although the mean of Tajima's *D* across loci is close to zero) argues against the simple explanation that the European population is in equilibrium. It is more likely that several different confounding processes have occurred during the habitat expansion of *D. melanogaster*, thus producing a mean value of *D* close to zero with a significantly higher-than-expected variance. Since some fragments show a significant haplotype structure (see RESULTS and online supplemental Table 2 at http://www.genetics.org/supplemental/), admixture following different colonization events may have shaped the observed pattern of polymorphism (in addition to the occurrence of a bottleneck). This scenario should lead to positive *D* values. The observed mean of Tajima's *D* of ~0 may therefore be explained by counteracting demographic and selective effects (*i.e.*, population size expansion following colonization and positive directional selection due to local adaptation, both producing negative *D* values).

Selection: The influence of demographic factors on the patterns of variation poses a problem for detecting possible footprints of selection. However, at least to some extent, this difficulty was overcome by our multilocus approach using a large number of fragments. As discussed above, it allowed us to get insights into demographic forces that shaped the standing variation in both populations. However, since the level of polymorphism across all fragments is on average relatively high, it was also possible to search for fragments with low variation that may be footprints of recent positive directional selection (selective sweeps).

In the highly variable African population, we did not find clear evidence for positive selection. Although we employed a series of neutrality tests (including the HKA test, Depaulis and Veuille's haplotype tests, and Fay and Wu's *H* test), only one test was significant in one frag-

ment. This observation is surprising. It may, however, not generally hold for African populations, as Mousset *et al.* (2003) found footprints of positive selection in a West African population.

Under a recurrent hitchhiking model, average Tajima's *D* value is expected to be negative due to a skew in the frequency spectrum toward an excess of rare variants (Braverman *et al.* 1995). We have observed this skew toward rare variants leading to an average negative Tajima's *D*. However, in contrast to Andolfatto and Przeworski (2001), who found a positive correlation between Tajima's *D* and recombination rates on a genome-wide scale (as expected under recurrent hitchhiking), we could not detect such a correlation on the X chromosome. The only signature of selection we observed in our sample was a (weak) correlation between recombination rate and levels of nucleotide diversity.

The data from the European population show two salient features: (i) a large number of fragments with zero or low levels of variation and (ii) a significant excess of derived variants at the low-variation loci (relative to the rest of the fragments) that are fixed in the European sample but rare in the African population. Both observations are difficult to explain without invoking positive natural selection. First, demographic modeling suggests that our observation of 13 fragments with zero variation is not consistent with a neutral equilibrium model or a neutral model of population founding followed by expansion. To explain our second finding, an evolutionary force needs to be postulated that brings newly arisen or rare African variants into high frequency in Europe in genomic regions of low variation (but not in the rest of the genome examined). It is difficult to imagine that any evolutionary force other than locus-specific positive directional selection is able to simultaneously produce both features i and ii. These results are consistent with the hypothesis that the European population has experienced frequent selective sweeps in the recent past during its adaptation to new habitats.

We thank K. Bhuiyan and H. Geisert for excellent technical assistance and M. Veuille and two anonymous reviewers for helpful comments on a previous version of this manuscript. This work was funded by the Deutsche Forschungsgemeinschaft (STE 325/6-1). S.M. was supported by a Marie-Curie Postdoctoral Fellowship from the European Union (MCFI-2002-01461).

LITERATURE CITED

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

Aquadro, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. Dev. **7:** 835–840.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and DNA polymorphism in *Drosophila*, pp. 46–55 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 C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

Begun, D. J., and C. F. Aquadro, 1995 Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140:** 1019–1032.

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

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.

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B Biol. Sci. **164:** 362–379.

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

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

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

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

Depaulis, F., S. Mousset and M. Veuille, 2001 Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol. Biol. Evol. **18:** 1136–1138.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Harr, B., M. Kauer and C. Schlötterer, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

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

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913–1931.

Lachaise, D., M. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup, pp. 159–225 in *Evolutionary Biology*, edited by M. K. Hecht, B. Wallace and G. T. Prance. Plenum, New York.

Lazzaro, B. P., and A. G. Clark, 2003 Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. Mol. Biol. Evol. **20:** 914–923.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.

Li, Y. J., Y. Satta and N. Takahata, 1999 Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes Genet. Syst. **74:** 117–127.

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

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

Mousset, S., L. Brazier, M.-L. Cariou, F. Chartois, F. Depaulis *et al.*, 2003 Evidence of a high rate of selective sweeps in African *Drosophila melanogaster*. Genetics **163:** 599–609.

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

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 *Drosophila simulans*. Mol. Biol. Evol. **18:** 291–298.

Robertson, L. S., 1975 Gene frequency distributions as a test for selective neutrality. Genetics **81:** 775–785.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

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

Communicating editor: M. Veuille