Patterns of Genetic Variation at a Chromosome 4 Locus of *Drosophila melanogaster* and *D. simulans*

Mark A. Jensen,[1] Brian Charlesworth[2] and Martin Kreitman

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637-1573

Manuscript received June 6, 2001
Accepted for publication October 23, 2001

ABSTRACT

DNA sequence surveys of *Drosophila melanogaster* populations show a strong positive correlation between the recombination rate experienced by a locus and its level of nucleotide polymorphism. In particular, surveys of the fourth chromosome gene ci^D show greatly reduced levels of nucleotide variation; this observation was originally interpreted in terms of selective sweeps occurring on the nonrecombining fourth chromosome. Subsequent theoretical work has, however, uncovered several other selective processes that can reduce variation. In this study, we revisit the Drosophila fourth chromosome, investigating variation in 5–6 kb of the gene *ankyrin* in *D. melanogaster* and *D. simulans*. Silent nucleotide site diversity is $\sim 5 \times 10^{-4}$ for both species, consistent with the previous observations of low variation at ci^D. Given the observed frequency spectra at *ankyrin*, coalescent simulations indicate that reduced diversity in the region is unlikely to be due to a selective sweep alone. We find evidence for recombinational exchange at this locus, and both species appear to be fixed for an insertion of the transposable element HB in an intron of *ankyrin*.

THERE is a strong positive correlation between the local recombination rate experienced by a locus and its level of nucleotide polymorphism in populations of *Drosophila melanogaster* (Aguadé *et al.* 1989; Begun and Aquadro 1992; Aguadé and Langley 1994; Aquadro *et al.* 1994; Moriyama and Powell 1996; Andolfatto and Przeworski 2001). This is one of the most striking patterns to have emerged from DNA sequence surveys. It persists after correction for possible differences in mutation rates among chromosomal regions by the use of interspecies sequence divergence data (Begun and Aquadro 1992; Aguadé *et al.* 1994). This pattern was initially taken to mean that neutral variants had been hitchhiked to fixation by strongly advantageous mutations to which they were closely linked (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1998, 2000; Gillespie 2000). The effects of such "selective sweeps" (Berry *et al.* 1991) are expected to be more pronounced in regions with low rates of recombination, because linkage between a given locus and a target of selection will on average be tighter in such regions (Aguadé *et al.* 1989; Begun and Aquadro 1992).

The relation between variability and recombination rate thus seemed to provide evidence for the frequent occurrence of adaptive gene substitutions throughout the Drosophila genome (Wiehe and Stephan 1993; Stephan 1995). But Charlesworth *et al.* (1993) identified a potentially widespread, nonadaptive, process that could result in reduction of variation in low-recombining regions. Their "background selection" hypothesis proposes that purifying selection against recurrent deleterious mutations also eliminates neutral variants at nucleotide sites closely linked to such mutations. Models of the effects of deleterious mutations on neutral variability across chromosome arms provide good fits to the observed relation between local recombination rate and the level of DNA sequence variation (Hudson and Kaplan 1995; Charlesworth 1996). For most regions of low diversity, therefore, background selection seems to be as likely a contributor to reduced diversity as hitchhiking due to favorable mutations. In addition, other processes are capable of reducing variation in regions of low recombination rates, including temporally fluctuating selection pressures (Gillespie 1994, 1997; Barton 2000) and Hill-Robertson interference effects among tightly linked, weakly selected variants (Comerón *et al.* 1999; McVean and Charlesworth 2000).

It is possible to get an idea of the extent to which selective sweeps may be important in contributing to the pattern just described by examining the frequency spectrum of neutral polymorphic sites in regions of reduced recombination. While this spectrum will be strongly skewed toward rare variants following a recent selective sweep (Braverman *et al.* 1995; Simonsen *et al.* 1995; Fu 1997; Fay and Wu 2000), the frequency spectrum associated with background selection or fluctuating selection is often not much perturbed from neutral expectation (Hudson and Kaplan 1994; Charlesworth *et al.* 1995; Gillespie 1997). A failure to detect a skewed frequency

[1]*Present address:* Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195-8070.

[2]*Corresponding author:* Institute of Cell, Animal and Population Biology, University of Edinburgh, W. Mains Rd., Edinburgh EH9 3JT, United Kingdom. E-mail: brian.charlesworth@ed.ac.uk

distribution may therefore suggest that selective sweeps are not the major factor in contributing to reduced variability. Previous work indeed suggests that significant departures from neutrality are detected only infrequently in regions of reduced recombination (Braverman *et al.* 1995; Charlesworth *et al.* 1995; Langley *et al.* 2000). However, a recent survey of published data suggests that there is usually a tendency toward a greater degree of skew toward low-frequency variants in regions of low recombination in African (but not non-African) populations of *D. melanogaster* (Andolfatto and Przeworski 2001), although individual loci do not show significant effects.

Given that there is little polymorphism in regions of low recombination, failure to detect departures from neutrality may simply reflect low power of the statistical tests. For this reason, it is important to gather more data on the properties of natural variation in regions of reduced recombination. In this study, we revisit the fourth chromosome of Drosophila. In *D. melanogaster,* this small (5–6 Mb) genetic element is not known to recombine under ordinary laboratory conditions (Sturtevant 1951; Hochman 1976) and is achiasmate at female meiosis (Hawley *et al.* 1993). It is thought to contain 74 genes (Adams *et al.* 2000), of which 12 have been characterized. Previous studies, performed on a relatively small scale, did not uncover enough polymorphism to be able to apply tests on the basis of the frequency spectrum. Berry *et al.* (1991) found no polymorphism among 10 *D. melanogaster* chromosomes and 1 singleton site among 9 *D. simulans* chromosomes for 331 silent sites of the ci^D gene, while Hilton *et al.* (1994) found no polymorphism over the same region among 4 *D. sechellia* chromosomes and a singleton site among 6 *D. mauritiana* chromosomes. Both of these studies interpreted the lack of polymorphism as evidence for a recent selective sweep, but the subsequent development of alternative models for explaining reduced variation casts doubt on this conclusion (see above).

We have examined 38 *D. melanogaster* lines and 33 *D. simulans* lines for ~5 kb of predominantly intronic sequence of the fourth chromosome gene *ankyrin* (Dubreuil and Yu 1994), to obtain better estimates of the levels of variation and frequency spectra on the fourth chromosome in these two species. To conduct this larger-scale study, we used the rapid mutation detection technology, denaturing high-performance liquid chromatography (DHPLC; Huber *et al.* 1993), to identify homologous DNA fragments that contain nucleotide or length variants. We have found that nucleotide site diversity is approximately the same in both species, 5×10^{-4}, a value similar to that found for silent sites in other regions of reduced recombination in *D. melanogaster* (Moriyama and Powell 1996; Langley *et al.* 2000). There is evidence for some recombinational exchange at this locus, based on the "four gamete test" of Hudson and Kaplan (1985). The presence of several intermediate-frequency variants at *ankyrin* in both species means that it is hard to explain the data on the basis of a selective sweep alone.

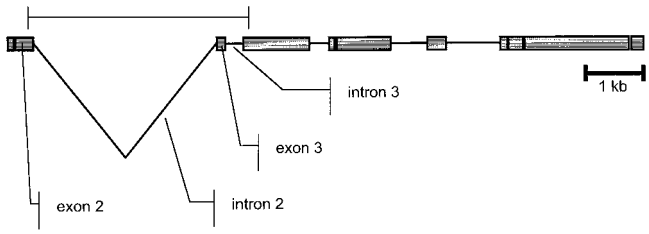

Figure 1.—Predicted genomic structure of *D. melanogaster ankyrin.* The structure is based on Berkeley Drosophila Genome Project accession no. CG1651. Shaded boxes indicate exons; lines indicate introns. Bar indicates region examined in this study.

MATERIALS AND METHODS

Drosophila stocks: We surveyed 38 isofemale lines of *D. melanogaster,* collected from three different localities. Eighteen lines were collected by M. Noor near Beltsville, Illinois, in 1991; 12 lines by C. Schloetterer near Beltsville, Illinois, in 1994; and 8 lines by M. Kreitman at Terhoon's Farm, Massachusetts, in 1989. For 32 of these lines, we extracted single fourth chromosomes using a *y;bw;*ci^D*/*ey^D marker stock. The extracted lines were made homozygous for either *yellow* or *brown,* to facilitate detection of contamination. No contamination was ever found in these homozygous lines. For the remaining 6, the extraction lines were lost, so that single flies from the original isofemale stocks were used in the PCR/DHPLC analyses described below. Thirty-three isofemale lines of *D. simulans* were surveyed from a single locality. These were originally collected by A. Berry, near Tempe, Arizona, in 1990. Single flies from each line were used for PCR/DHPLC. Stocks were maintained at 18°, on standard yeast-sucrose-cornmeal-agar medium.

Gene region: We analyzed portions of exons 2 and 4, the whole of exon 3, and the second and third introns of the *ankyrin* gene in both species (Figure 1). *Ankyrin* was localized to chromosome 4 of *D. melanogaster* by Dubreuil and Yu (1994), using a cDNA probe. The introns were identified during the course of PCR amplification of *D. melanogaster* genomic DNA, using a set of primer pairs designed from the *ankyrin* cDNA sequence (GenBank accession no. L35601). A primer pair (ANK+38, ANK−275; see below) near the 5′ end of the cDNA did not amplify, using a typical protocol designed to amplify 1 kb; we used this pair in a long PCR reaction (Expand Long PCR; Roche Molecular Biochemicals, Indianapolis) and obtained a product of ~6 kb. Upon sequencing (see below), this proved to contain one long (5.4-kb) intron (*ankin2*), a 108-bp exon (*exon3*), and a 363-bp intron (*ankin3*). The complete genomic structure of *ankyrin* was determined by the Drosophila genome project (gene CG1651; Adams *et al.* 2000) and confirms this structure, except that our annotation compared to the genome project annotation suggests an additional small intron upstream of our region. A correction has been forwarded to Flybase.

S. Assimacopoulos kindly verified the chromosomal location of the intron by *in situ* hybridization to polytene chromosomes, using a *Hin*dIII restriction fragment from the long PCR product. The same primer pair amplified the homologous region in *D. simulans.* This product was a 5-kb fragment, with structure

similar to *D. melanogaster*; in particular, there is one long and one short intron, and *exon2* is the same length in the two species.

DNA sequencing: To sequence the *D. melanogaster* region, a shotgun sequencing protocol developed by P. Andolfatto was used. Briefly, the 6-kb *D. melanogaster* product was mechanically sheared, polished, and blunt-end cloned into a derivative of vector pZero (Invitrogen, San Diego). Approximately 30 positive clones were picked for colony PCR, using universal (M13-20 and M13rev) primers. The resulting products were dye-terminator cycle sequenced using a premixed reaction (PE Biosystems/ABI) and sequenced on an automated sequencer (ABI 377). All portions of the region were sequenced at least twice. The same protocol was used on the *D. simulans* 5-kb product. This yielded four long contigs. To finish the sequencing, the 5-kb product was cut with a six-base blunt-cutting restriction enzyme (*Dra*I) that cut rarely among the contigs. Subcloning and sequencing the resulting fragments provided the missing *D. simulans* sequence. Ultimately, we obtained 6071 bp of *D. melanogaster* sequence and 5057 bp of *D. simulans* sequence (GenBank accession nos. AY054998 and AY054997, respectively). The sequence for *D. melanogaster* is from line B45 (Terhoon's Farm population). The sequence for *D. simulans* was obtained from line s52 genomic DNA for the initial shotgun sequencing, and a cloned long PCR product from line s18 was used to fill contig gaps. The nucleotide coordinates used below are such that +1 indicates the first base of intron 2 in our notation.

PCR/DHPLC: We used DHPLC of DNA fragments (Huber *et al.* 1993) to identify nucleotide and length differences between homologous fragments of *ankin2-3* among lines. We designed primer pairs from the intron sequence to amplify 150- to 400-bp fragments that together span the entire region (Jensen 2000). The primer pairs did not perfectly overlap, but no gap was >105 bp for either species. The pairs covered 5296 usable base pairs (*i.e.*, excluding actual gaps between fragments and the primer sequences) in *D. melanogaster* and 4459 usable base pairs in *D. simulans*.

Template DNA for the fragment PCRs was generated as follows. Genomic DNA was extracted from a single fly for each isofemale or chromosome-extracted line. This DNA was used as template for a long PCR reaction, using a high-fidelity DNA polymerase (Expand Hi-Fidelity; Roche Molecular Biochemicals). For *D. melanogaster* lines, the entire *ankin2-3* region was amplified using primers ANK+38 (5′-CGCTTGGTGATGTAC GAGTTG-3′) and ANK−275 (5′-TGTCCACATATCCGTCCT TTG-3′). For the *D. simulans* lines, only the *ankin2* intron was amplified, using *simulans*-specific primers designed from exons 2 and 3 (sankin1U57, 5′-TAATGGAATGGCTTTAGACA ACAA-3′; and sankin1L169, 5′-TATGTCCGATATTTCTCCA CAGTC-3′), since not every line would amplify well using the *melanogaster* primers. This long PCR product was used directly as template for the fragment PCRs for all *D. melanogaster* lines and 23 of the *D. simulans* lines. The long PCR product for the remaining *D. simulans* lines was cloned into the TOPO-4 vector (Invitrogen), according to the manufacturer's protocol. For these lines, we used cloned *ankin1* from a plasmid prep as template for the fragment PCRs.

Using PCR product as template for secondary PCR reactions raises the issue of "PCR error," or the occasional incorporation of noncomplementary bases by the thermostable DNA polymerase during PCR amplification, which may result in artifactual variants. This is of particular importance in the study of regions of very low variation like the fourth chromosome. For a high-fidelity polymerase, such as that used to produce *ankin1* template here, there is little cause for concern if the PCR product from genomic DNA is used directly as template (for a detailed analysis, see Jensen 2000). But when cloned PCR product is used as template for a line, there is a distinct possibility that at least one base along its length has been changed by misincorporation. Such spurious variants should always appear as singletons and can thus be identified by checking two different clones derived from the same PCR product. In this study, three singleton variants were found among cloned *D. simulans* lines, of which two were rejected and one accepted.

DHPLC was used to survey fragments for variants as follows. For each fragment or primer pair, all lines were amplified using ordinary *Taq* polymerase (QIAGEN, Valencia, CA) in a 25-μl reaction, using 1 μl of 1/100 dilution of the template PCR described above (details are given in Jensen 2000). One line was chosen as a comparison standard and was amplified to provide 8 μl standard reaction per line. All fragment PCRs were performed with identical primer concentrations and were primer limited. Thus, final product concentrations were approximately equal for all lines. To prepare samples for DHPLC analysis, 8 μl of each PCR reaction was mixed with 8 μl of product from the standard line, heated to 98° for 3 min to denature, and cooled to 65° over 30 min in a thermal cycler. Five microliters of this denatured and reannealed product was passed over the DHPLC column in a solvent gradient and at a temperature determined from the melting characteristics of the fragment. These parameters were determined by using WAVEMaker software (Transgenomic Inc.). UV-absorbance chromatographs for each line were compared with a chromatograph of the unmixed standard reaction that had been subjected to denaturation/reannealing. Chromatographs reveal variants, because their shapes change with the presence of heteroduplex DNA in the sample. If a line contains a variant with respect to the standard line, the mixed PCR reaction will yield approximately one-half heteroduplex DNA, which under appropriate conditions will give a chromatograph that is different in shape from the unmixed standard. If the standard line and the tested line are identical in sequence, most of the DNA will be homoduplex, and the chromatograph will mimic the standard. As this study shows, the technique can unambiguously reveal single base pair substitutions, single base pair insertion-deletions (indels), and longer indels.

When variants appeared among the lines for a fragment, on the basis of a subjective observation of associated chromatograms, lines were assigned a DHPLC variant type number according to the chromatogram shape. When ambiguities arose in chromatogram interpretation, more types were assigned. Generally, slight differences between chromatograms, such as small changes in retention times of absorbance peaks, did not indicate an underlying mutation. Gross changes in chromatograms between fragments, which resulted in differences in numbers of peaks or marked width differences in single peaks, were reliable indicators of underlying sequence differences. In two instances, variant classes with gross chromatogram differences were nevertheless isosequential. It is likely that nonspecific amplification in the PCR reactions led to these results; we did not follow up these cases. Ultimately, only gross differences were scored as separate DHPLC variants. At least three lines were sequenced, if possible, within each DHPLC variant class, including the standard class, using the original fragment PCR reaction as template in a cycle-sequencing reaction. Unsequenced lines within a DHPLC variant class were assumed to contain the same sequence variants as the sequenced members of that class. Most of those fragments whose chromatographs appeared to be the same as the standard for every line were assumed to be monomorphic and were not analyzed further. This is justified by our preliminary data and the fact that no sequence variation was ever found within DHPLC variant classes in polymorphic fragments. Details are given in Jensen (2000).

It is possible that this survey will not have identified all polymorphisms in the region for these lines. We attempted to mini-

mize the possibility of missing variation by performing DHPLC at multiple column temperatures, when the fragment was predicted to contain several melting regimes (*i.e.*, heterogeneity in GC content); this has been shown to increase the chances of detecting point variation within high-melting-temperature tracts (Transgenomic Inc., personal communication). Representatives of DHPLC classes were sequenced to characterize the underlying sequence changes and to demonstrate the reproducibility of the chromatogram-sequence association within classes. However, it is unlikely that failure to identify a variant would depend on its population frequency, so that the broad conclusions from this study should be little affected by DHPLC inefficiency. On the other hand, there should be no spurious sequence variation introduced by the survey method, since DHPLC variant classes were conservatively assigned and checked by direct sequencing. In fact, this source of error should be reduced relative to direct sequencing alone, since DHPLC and direct sequencing provide independent ways of checking sequence identity.

Evolutionary parameter estimates: Estimates of evolutionary parameters, including θ_w [Watterson's (1975) estimator of the scaled mutation rate $\theta = 4N_e u$], the nucleotide site diversity π estimator of θ (the mean pairwise difference per base pair; Tajima 1983), and Tajima's (1989) D statistic for measuring departure of the site frequency spectrum from neutrality, were calculated using either DNAsp (Rozas and Rozas 1997) or SITES (Hey and Wakeley 1997). Estimates of the scaled recombination rate $C = 4N_e c$ (Hudson 1987) and F_{ST} (Wright 1951) for the *D. melanogaster* data were calculated with SITES. In addition, Dr. Jeffrey Wall kindly calculated the scaled rate of gene conversion, $G = 4N_e g$, on the assumption that all intragenic recombination is caused by gene conversion (Frisse *et al.* 2001). Here, N_e is the effective population size, u is the mutation rate per nucleotide site, and c and g are the rates of crossing over and gene conversion per nucleotide, respectively, averaged over males and females.

***D. melanogaster-D. simulans* sequence alignment:** Relatively long regions of homology are present between the two species, but the region has been subject to many large insertion/deletion events since divergence between the species. This makes the results of typical alignment algorithms unreliable. However, we wished to avoid as much subjective alignment as possible, to get a reasonably unbiased estimate of divergence. We combined alignment by dotplot and the Needleman-Wunsch algorithm (Weir 1996, Ch. 9), as implemented in the MegAlign program (DNAStar) as follows. We started with a sliding alignment window of 50 bp (chosen to eliminate signals from microsatellite repeats). A window alignment was rejected if the percentage base identity was significantly low in the following sense. The average divergence for noncoding regions between *D. melanogaster* and *D. simulans* is 0.061 (Moriyama and Powell 1996), so that a 50-bp region should contain three divergent sites on average. Assuming a Poisson distribution of divergent sites, the probability of eight or more divergent sites in 50 bp is <0.025.

On the basis of this procedure, the dotplot parameters were set to accept a window alignment with 86% identity or better. This resulted in the dotplot in Figure 2; the figure changes little with changes of a few percent similarity in either direction. Each aligned diagonal was passed to the Needleman-Wunsch algorithm, to assign gaps objectively. Excluding gaps, 2195 bp were aligned by this protocol. Further subalignments were performed in the analysis of the HB element insertion (see results).

Simulations: Since it has been suggested that the fourth chromosome of *D. melanogaster* has undergone a recent selective sweep (Berry *et al.* 1991; Hilton *et al.* 1994), we ask whether our more extensive data set is also compatible with a selective sweep model. Many scenarios involving both strong and weak selection, which include partial sweeps, multiple sweeps, or sweeps with significant recombination, can be devised to explain this or any other data set (Gillespie 1997, 2000; Fay and Wu 2000; McVean and Charlesworth 2000). But if we consider only a simple model, we can use simulation techniques to explore the relative likelihoods of the data in a specified context and perhaps restrict the number of possible explanations of the data (see discussion).

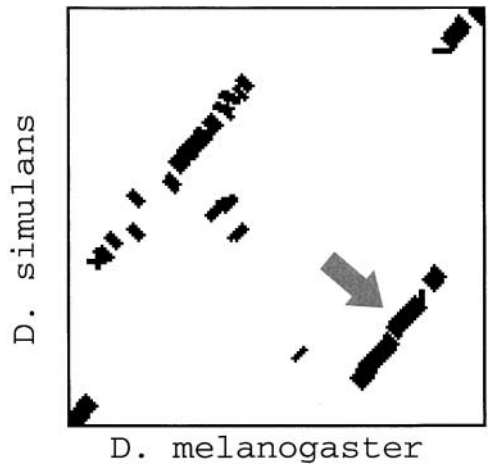

Figure 2.—Dotplot alignment of *D. melanogaster* and *D. simulans ankyrin* region. Arrow indicates HB-like transposable element (see text for details).

In these simulations, we consider only "catastrophic sweeps" (Perlitz and Stephan 1997) under the following implicit assumptions: (A) Selection is strong relative to mutation, such that no variant arising during the process of fixation of the selected allele is likely to be sampled; (B) selection is strong relative to recombination, such that no sampled allele is likely to have undergone a recombination event during the fixation of the selected allele. Assumption A is quite weak; assumption B is reasonable for the fourth chromosome, despite the evidence of genetic exchange in the data set, since the absence of crossing over on this chromosome suggests that this exchange results mainly from gene conversion, which causes only low levels of recombination over intervals longer than typical genes (see discussion).

Under these assumptions, a selective sweep completely eliminates variation in a given region. If it is further assumed that evolution proceeds neutrally following the sweep under an infinite-sites model, coalescent techniques could be used to simulate samples drawn from the population. Assume a given time T_s, since the last selective sweep (in units of $2N_e$ generations), a scaled mutation rate $\theta = 4N_e u$, and a sample of n alleles. Coalescent events can be simulated according to the standard neutral model with constant population size (Hudson 1990), until the accumulated time is $>T_s$, at which point the value T_s is assigned to subsequent nodes. This effectively truncates the coalescent at the time of the sweep, creating a star-shaped genealogy. Mutations are added to each branch by sampling from a Poisson distribution with mean $(\theta t/2)$, where t is the length of the branch in units of $2N_e$ generations.

The simulations can be used as follows to estimate the likelihood of the pair (S, K), the observed number of segregating sites, and average pairwise difference between alleles for a given pair of sweep parameters (θ, T_s). Assume that B samples are generated for each pair of sweep parameters. Write

$$M_\delta = \sum_{j=1}^{B} I_\delta(S, K, j), \qquad (1)$$

TABLE 1

***Ankyrin* polymorphism statistics**

	No. sites	θ_w	π	D
		D. melanogaster		
Biallelic nucleotide	10	2.38/0.45[a]	2.75/0.52	0.47
Indels	4	0.95/0.18	1.20/0.23	0.62
All sites	17	4.05/0.76	4.21/0.79	−0.06
		D. simulans		
All biallelic nucleotides	9	2.22/0.50	1.72/0.39	−0.68
Replacement	1	0.25/1.43[b]	0.06/0.36	
Synonymous	1	0.49/10.96[c]	0.12/2.61	
Indels	4	0.99/0.22	1.58/0.35	1.51
All sites	13	3.20/0.72	3.30/0.74	0.10

[a] Value before slash is estimate for region; per-base pair value $\times 10^3$ follows slash.
[b] Number of replacement sites: 172.
[c] Number of synonymous sites: 44.

where for the jth iteration,

$$I_\delta(S, K, j) = \begin{cases} 1, & \text{if } |K_j - K| \le \delta \text{ and } S_j = S \\ 0, & \text{otherwise.} \end{cases}$$

Here, δ is a preassigned mesh size for the continuous variable K, and S_j and K_j are the simulated number of segregating sites and mean pairwise difference between alleles for the jth replicate. Following Weiss and von Haeseler (1998), the likelihood of the sample is approximated by

$$L(S, K, \mid \theta, T_s) = \frac{1}{B} M_\delta(S, K). \qquad (2)$$

In this study, $B = 10^5$ and $\delta = 0.1$, where δ was chosen to give a fairly smooth likelihood surface. Graphical rendering of the likelihoods was performed with Mathematica 3.0 (Wolfram Research, Champaign, IL).

While it is assumed for the purposes of simulation that recombination is unlikely to have occurred during the substitution of a strongly selected allele, recombination in the genealogy cannot be excluded altogether, on the basis of the evidence contained in the data for both species (see RESULTS). Recombination following a sweep is likely to make significance tests based on the joint distribution of S and K under the above model conservative, since recombination is known to reduce the variance of the distributions of S and K (Hudson 1990; Wall 1999).

RESULTS

Polymorphism data: The estimates of the levels of nucleotide polymorphism per site, θ, displayed in Table 1, are significantly reduced compared to the published genome-wide averages for *D. melanogaster* (noncoding average, 0.01) and *D. simulans* (noncoding average, 0.02; Moriyama and Powell 1996); the significance level ($P < 2 \times 10^{-5}$) was established by neutral coalescent simulations. Pairwise F_{ST} values were estimated for the *D. melanogaster* populations using single nucleotide polymorphisms; the largest value was 0.13. We therefore treated the entire data set as a sample from a single panmictic population.

The identities of the polymorphic sites are displayed in Tables 2 and 3. *D. melanogaster* polymorphisms were found only in intron 2. *D. simulans* also had polymorphic nucleotide sites in exon 3 and intron 3; one replacement variant and one silent variant were observed in exon 3, and both are low-frequency sites. *D. simulans* had 9 biallelic single-nucleotide variants; *D. melanogaster* had 9 biallelic sites and one tri-allelic site. For the purposes of the simulations, the double-hit site was treated as two singleton sites. Both species also have single-base and short indels segregating. There are 9 indel variants out of 30 total variants for both species pooled. This is not significantly different from the value of 49 indels out of 191 total polymorphisms in noncoding sequences from the *su(s)* and *su(w^a)* regions in *D. melanogaster* (Langley *et al.* 2000) or the value of 11 indels out of 59 total differences between *D. melanogaster* and *D. simulans* for the *Lcp*ψ pseudogene (Pritchard and Schaeffer 1997). It therefore seems that indels in relatively unconstrained regions make up ~25% of all variants in Drosophila, which is substantially higher than the ~10% reported for humans (Nachman and Crowell 2000). The intron sequences are AT rich (64 and 69% AT for introns 2 and 3 in *D. melanogaster*, and 66 and 71% in *D. simulans*). Repeats were screened for, using a standard program (Benson 1999). In *D. melanogaster*, TGAAAAGTA is repeated five times at 1890–1936 and AAGTATGAAAAGCATTTGAA is repeated twice at 1906–1946. There are no tandem repeats in the *D. simulans* intron sequence, and none of the indel variants are associated with the tandem repeats.

No site is polymorphic in both species. Table 4 indicates polymorphic sites that have an identifiably homologous site in the sister species. For each such site, the inferred state of the rare variant is derived. These results are used in relation to the problem of analyzing a sweep with recombination (see DISCUSSION). The data are too sparse to determine whether deletions and insertions differ in their frequencies of occurrence; an excess of

TABLE 2

***D. melanogaster* polymorphic sites**

			7	8	1 1	2 0	3 6	3 7	3 9	4 1	4 6	4 7	5 1	5 1	5 2	5 3	5 3
	2	4	5	9	0	2	2	1	9	0	1	8	3	9	7	6	7
	4[a]	8	9	6	9	3	9	5	2	7	2	6	4	5	2	6	5
Cons[b]	TA	T	C	A	G	G	−	−	C	T	C	−	G	C	−	T	−
B26	.	.	.	.	.	.	.	.	.	.	A	.	T	.	.	.	.
B49	.	A	T	.	.	.	.	.	.	.	.	.	.	T	d	.	.
I39	.	.	.	.	.	.	.	.	.	.	A	.	T	.	d	.	i[c]
I44	.	.	.	.	.	.	.	.	.	.	.	D[d]	.	.	d	.	.
I80	.	.	.	.	.	.	.	.	.	.	A	.	T	.	d	.	.
T4	.	.	.	.	.	.	.	.	.	.	A	.	T	.	d	.	.
T41	.	.	.	.	.	.	.	.	.	.	A	.	.	.	d	.	.
T44	.	A	.	G	.	.	d[e]	.	.	.	.	.	.	.	.	A	.
B32	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
T22	.	.	.	.	T	A	.	.	.	.	A	.	T	.	d	.	.
B17	.	.	.	.	.	.	.	.	.	.	A	D	.	.	d	.	.
B43	.	.	.	.	.	.	.	.	.	.	A	.	.	.	d	.	.
B2	.	.	.	.	.	.	.	.	.	.	A	D	.	.	d	.	.
T25	.	.	.	.	A	.	d	.	G	A	.	.	.	.	.	.	.
B6	.	A	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B13	.	A	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B44	.	A	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B31	.	A	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B42	.	A	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
I37	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B12	.	A	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.
B50	.	A	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.
I66	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
B23	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
I34	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
I88	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
I94	.	.	.	.	.	.	.	.	.	.	A	D	.	.	d	.	.
I101	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
T24	.	.	.	.	.	.	.	M[f]	.	.	A	.	T	.	d	.	.
T48	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
B15	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
I10	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
I43	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
B11	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
T40	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	.	.
I97	AT	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
B9	AT	.	.	G	.	.	d	.	.	.	.	.	.	T	.	.	.
B4	.	.	.	G	.	.	d	.	.	.	.	.	.	.	.	A	.
[g]Freq	2	9	3	16	□	1	17	1	1	1	11	4	6	10	12	14	1

Line identifiers are in left column. Underlines identify lines for which isofemale cultures were sampled; the rest were fourth chromosome homozygous.

[a] First intron 2 base = +1.

[b] Consensus base in this row.

[c] Single-base insertion (relative to consensus).

[d] 26-bp deletion (relative to consensus).

[e] Single-base deletion.

[f] Complex mutation: TAA to AAAA.

[g] □, three-allele site.

deletions has been reported in previous studies (Comerón and Kreitman 2000).

For 3 isofemale *D. simulans* lines out of the 39 that were initially scanned (Tempe, Arizona lines 32, 52, and 70), long PCR using gDNA template and *D. melanogaster* primers ANK+38 and ANK−275 amplified a product that approached 10 kb in length. Presumably this represented a transposable element insertion. Since it was

TABLE 3

***D. simulans* polymorphic sites**

	1	1	2	2	2	2	4	4	4	4	4		
	2	7	8	9	6	8	9	9	2	3	5	5	8
	8	3	6	6	8	0	2	5	9	3	4	6	6
	3	8	5	7	8	1	3	8	9	9	1	9	5
Cons	−	T	G	A	T	T	T	−	T	−	T	T	−
s4	.	.	.	G	G	.	.	.	.	.	.	.	D[a]
s6	.	.	.	G	G	.	.	.	.	.	.	.	.
s8	.	.	.	G	G	.	.	.	.	.	.	.	.
s10	.	.	.	G	G	.	.	.	.	.	.	.	.
s28	.	.	.	G	G	.	.	.	.	.	.	.	.
s37	.	.	.	G	G	.	.	.	.	.	.	.	.
s38	.	.	.	G	G	.	.	.	.	.	.	.	.
s45	.	.	.	G	G	.	.	.	.	.	.	.	.
s53	.	.	.	G	G	.	.	.	.	.	.	.	.
s62	.	.	.	G	G	.	.	.	.	.	.	.	.
s20c	.	.	.	G	G	.	.	.	.	.	.	.	.
s49c	i[b]	.	.	.	.	.	.	.	.	.	.	.	D
s57	i	.	.	.	.	.	.	.	.	.	.	.	D
s23c	i	.	.	.	.	.	.	.	.	I[c]	.	.	.
s26c	i	.	.	.	.	.	.	.	.	I	.	.	.
s33	i	.	.	.	.	.	.	.	.	I	.	.	D
s34	i	.	.	.	.	.	.	.	.	I	.	.	D
s44	i	.	.	.	.	.	.	.	.	I	.	.	D
s58	i	.	.	.	.	.	.	.	.	I	.	.	D
s64	i	.	.	.	.	.	.	.	.	I	.	.	D
s71	i	.	.	.	.	.	.	.	.	I	.	.	D
s82	i	.	.	.	.	.	.	.	.	I	.	.	D
s83	i	.	.	.	.	.	.	.	.	I	.	.	D
s79	i	.	.	.	.	.	.	.	.	I	.	A	D
s39	i	.	.	.	.	.	.	.	.	I	.	A	D
s14	i	.	.	G	G	.	.	.	C	I	.	.	D
s12	.	.	.	.	.	.	.	.	C	.	.	.	.
s48c	.	.	.	.	.	.	.	.	C	.	.	.	.
s18c	.	.	.	.	.	.	.	.	C	.	C	.	.
s24c	.	.	.	.	.	.	.	.	C	.	.	.	.
s31c	.	.	.	.	.	C	.	.	C	.	.	.	.
s80c	.	.	A	.	.	.	.	.	C	.	.	.	.
s40c	.	C	.	.	.	.	C	d[d]	.	I	.	.	.
Freq	15	1	1	12	12	1	1	1	7	14	1	2	14
[e]Type	n	n	n	n	n	n	n	n	n	n	r	s	n

See Table 2 legend. All *D. simulans* lines were isofemale without special effort made to extract fourth chromosomes. Lines labeled "c" were analyzed using cloned long PCR products, rather than single-fly genomic DNA, as template.

[a] 11-bp deletion.
[b] Single-base insertion.
[c] 20-bp insertion.
[d] Single-base deletion.
[e] Site type: n, noncoding; r, replacement; s, synonymous.

quite rare, we chose not to examine these lines further. No such insertion polymorphism was present in the *D. melanogaster* lines.

Between-species sequence comparisons: Figure 2 shows the dotplot alignment between the two species of the entire sequenced region. The overall divergence in the ~2 kb of alignable sequence is 12%. However, 1.2 kb of this involves an insertion into intron 2 of both species of a transposable element with high homology to the HB element of *D. melanogaster*; the element is indicated on the dotplot.

HB is a little-studied member of the P (Drosophila)/Tc (*Caenorhabditis elegans*) family of transposons and is characterized by a single open reading frame (ORF), flanked by direct repeats and additional DNA and bounded by short terminal inverted repeats (TIRs). The insertions are in-

TABLE 4

Alignable polymorphic sites

Site coordinate (consensus)	Homolog coordinate (identity)	Rare variant/ polarity	Freq
	D. melanogaster: six sites		
4612 (C)	1027[a] (C)	A/derived	11
4786[b]	1200[a]	Δ(26 bp)/derived[c]	4
5134 (G)	1610[a] (G)	T/derived	6
5195 (C)	1666[a] (C)	T/derived	10
5272	1744	Δ(T)/derived	12
5375 (8T)	4484 (8T)	∇(T)/derived	1
	D. simulans: five sites		
738 (T)	4378[a] (T)	C/derived	1
2801 (T)	1350 (T)	C/derived	1
4541 (T)	5433 (T)	C/derived	1
4569 (T)	5460 (T)	A/derived	2
4865	5845	Δ(11 bp)/derived	14

[a] Homolog found within HB-like element insertion.
[b] Entries without consensus do not exhibit the insertion or deletion corresponding to the rare variant.
[c] Δ indicates deletion, ∇ indicates insertion.

verted relative to *ankyrin* transcription (*i.e.*, the ORF and *ankyrin* would be transcribed in opposite directions) and are relatively closer to exon 2 in *D. simulans* than *D. melanogaster*. In *D. melanogaster*, HB is inserted into positions 4009–5301, corresponding to positions 226–1635 of the standard sequence of HB (GenBank accession no. X01748). In *D. simulans*, HB is inserted into positions 182–1774, corresponding to bases 201–1867 of the standard HB sequence.

The variation surveys show that the element is fixed within both species, although it occupies different locations. Nonhomologous DNA flanks both insertions. The divergence between the two elements alone is ~13%, while divergence for homologous DNA excluding HB is ~10%. Tables 5 and 6 give the pairwise divergence and insertion/deletion data in comparisons with the *D. melanogaster* HB standard GenBank sequence. The interpretation of these results is considered in the DISCUSSION.

TABLE 5

Divergences (excluding gaps) among HB-like elements

	Mel	Sim	Mel × sim
Common[a]	0.137	0.122	0.138
ORF[b]	0.108	0.105	0.128
Full seq[c]		0.127	
Ex-HB[d]			0.097

Columns are *D. melanogaster* (Mel) aligned with HB, *D. simulans* (Sim) with HB, and *D. melanogaster* with *D. simulans* (Mel × Sim) HB-like regions.
[a] HB-like sequence common to both species (1208 bp).
[b] HB open reading frame (GenBank HB coordinates 575–1021) only; *melanogaster* contains only 415 bp of the ORF.
[c] *D. simulans* insertion aligned to entire HB element (1572 bp).
[d] *D. melanogaster* and *D. simulans* homologous sequence outside HB-like element (987 bp).

Recombination estimates: Using the four-gamete test of HUDSON and KAPLAN (1985), we can infer that at least three recombination events have occurred in both the *D. melanogaster* and *D. simulans* lineages. Haplotype networks drawn by the method of BANDELT *et al.* (1999) for the two species are shown in Figures 3 and 4, indicating the positions of homoplasies that are likely to be caused by recombination events. Because the level of polymorphism is so low, it is unlikely that any of the commonly used estimators of scaled recombination rate would give an accurate estimate of the actual parameter (WALL 2000). Hudson's C estimator (HUDSON 1987) is $\sim 6.0 \times 10^{-3}$/bp for both species. The ratio of C to θ for *ankyrin* is ~10, rather higher than the values typically reported for these species (ANDOLFATTO and PRZEWORSKI 2000), but this may simply reflect the high error variance of the estimates of C. The corresponding estimates of the scaled rate of gene conversion for the entire region, assuming an exponential distribution of conversion tract lengths with a mean of 350 (J. WALL, personal communication), were 5.3×10^{-3}/bp and 6.8×10^{-3}/bp for *D. melanogaster* and *D. simulans*, respectively. The ratios of G to θ are ~10; with a mutation rate of $\sim 2 \times 10^{-9}$ per nucleotide (KEIGHTLEY and EYRE-WALKER 2000), the rate of gene conversion per site in female meiosis is estimated to be $\sim 4 \times 10^{-8}$/bp. This is two orders of magnitude lower than the value determined experimentally for the *rosy* locus (HILLIKER and CHOVNICK 1981), suggesting that the rate of gene conversion may be much lower for *ankyrin*. This should be qualified, however, by the fact that population-based estimates of rates of exchange for other loci in these species are also generally much lower than expected on

TABLE 6

HB-like element degeneration

	HB coord	Length (bp)	Mel	Sim	Feature[a]
Deletions	189–200	12	N/A[b]	+	TIR
	331–336	6	−	+	
	418–472	55	+	−	
	518–554	37	−	M[c]	
	555–568	14	+	−	
	592–595	4	+	−	ORF
	995–1038	44	+	−	ORF
	1236–1240	5	+	−	
	1480–1485	6	+	−	
Insertions	649	6	+	−	ORF
	1838	4	N/A	+	TIR

Insertions and deletions are relative to the *D. melanogaster* HB transposon. Coordinates are from GenBank accession no. X01748.

[a] TIR, event occurs within the HB terminal inverted repeats; ORF, event occurs within the HB open reading frame.

[b] *D. melanogaster* does not possess homologous HB sequence.

[c] Deleted region replaced with 6-bp nonhomologous DNA.

the basis of laboratory measurements of recombination (Andolfatto and Przeworski 2000).

Simulation results: The results of the simulations of catastrophic sweeps (see materials and methods) are summarized in Figures 5 and 6. All the single nucleotide polymorphisms were used for the observed results. The shading in the log-likelihood plots indicates their differences of the log-likelihoods of the observed S and K values from the maximum log-likelihood found in the simulations. Each cell corresponds to 50,000 iterates of the modified coalescent, using the underlying θ and T_s indicated on the axes. It can be seen that likelihoods

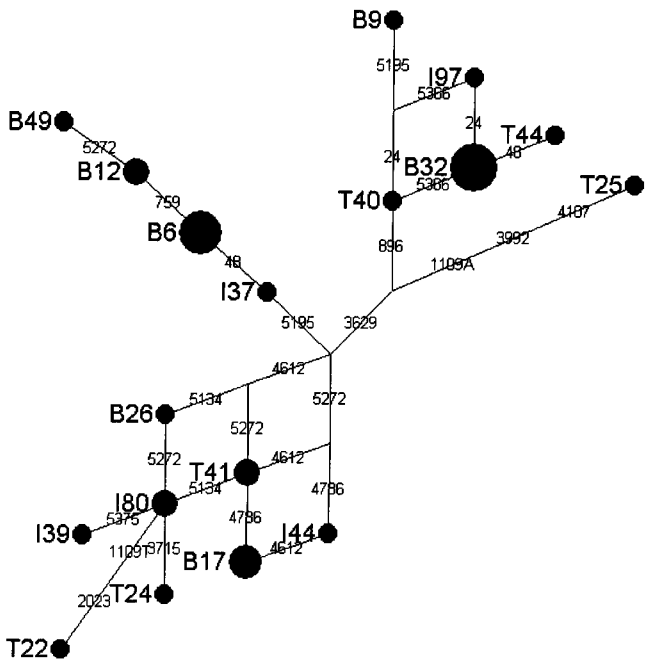

Figure 3.—Median-joining haplotype network for the observed *D. melanogaster* haplotypes. Reticulations arise due to homoplasies that are likely to have been generated by genetic exchange. Nodes are labeled with a strain name possessing the haplotype; node sizes are proportional to the haplotype frequency. Mutating sites are noted along the branches. The network was calculated and rendered with Network3.015 (available at http://www.fluxus-engineering.com; Bandelt *et al.* 1999).

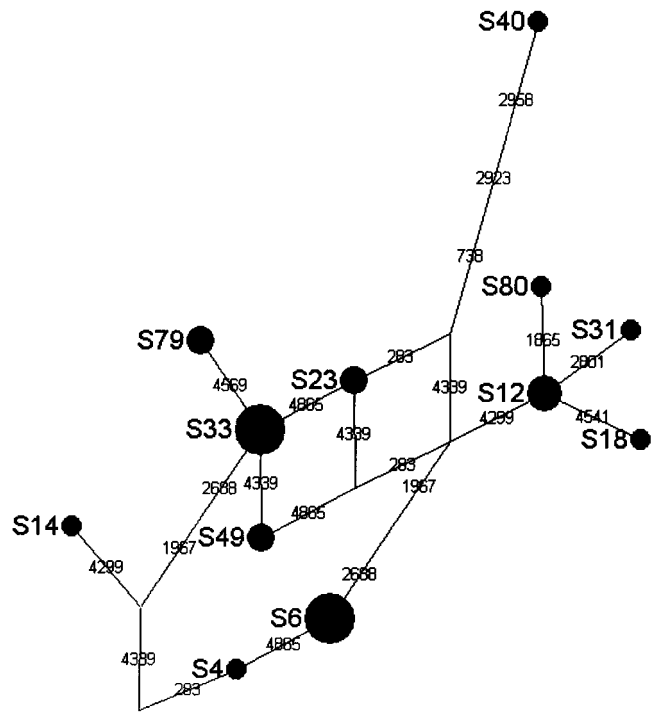

Figure 4.—Median-joining haplotype network for the observed *D. simulans* haplotypes.

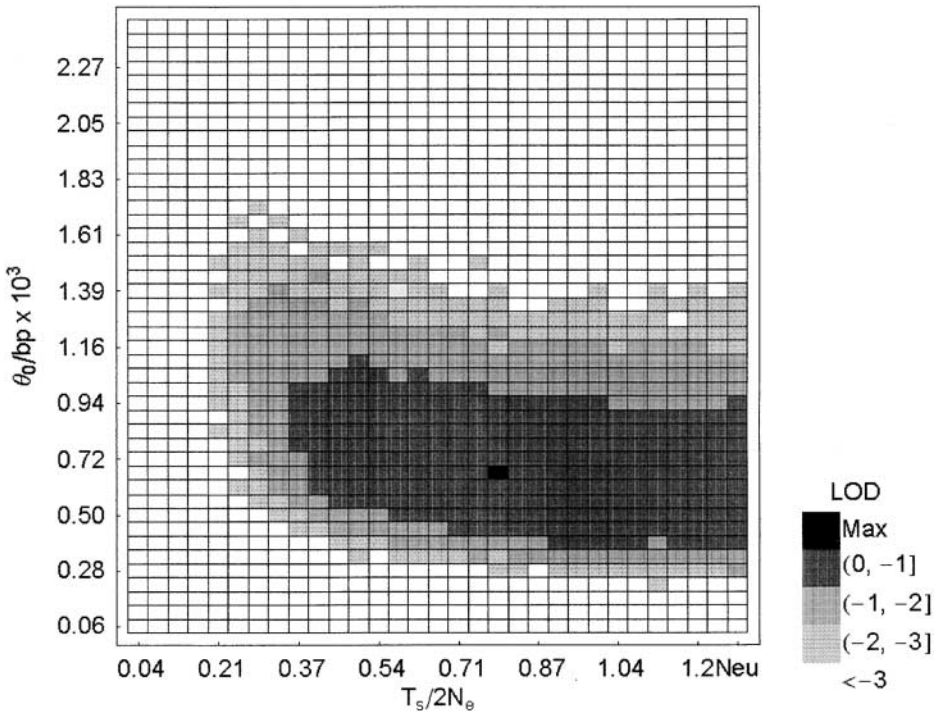

FIGURE 5.—Likelihood plot, *D. melanogaster ankyrin* intron simulation. Observed data are $S = 10$, $K = 2.75$, $n = 38$. Contours are shaded according to log-likelihood relative to the maximum (black-shaded cell). See text for details.

within 2 or 3 support units are found only for very low values of the underlying θ and relatively large values of the time T_s since the assumed sweep. The results show that there is a band of probable θ values that is relatively unchanging with possible sweep times and that the genome-wide average θ values of the order of 1% (see above) for noncoding sites are well outside this band for both species. In other words, the data indicate that, even under the assumption of a recent sweep, the underlying equilibrium diversity for the *ankyrin* region must be much lower than the standard value for silent sites, so that some other force or forces must have acted to reduce variation. In addition, the simulations allow the most recent sweep times to be rejected at the 5% level or better.

DISCUSSION

Reduced variability at *ankyrin*: Our results are consistent with the previous findings of low sequence variabil-

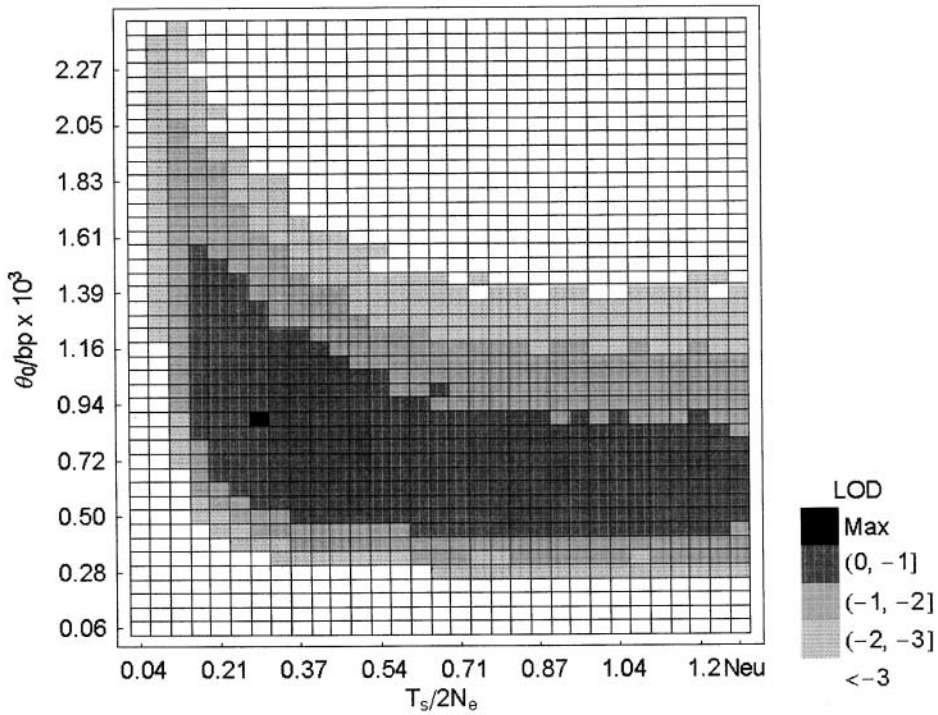

FIGURE 6.—Likelihood plot, *D. simulans ankyrin* intron simulation. Observed data are $S = 9$, $K = 1.72$, $n = 33$. Shading as in Figure 5.

ity at the chromosome 4 locus ci^D in *D. melanogaster* and its close relatives (Berry *et al.* 1991; Hilton *et al.* 1994). Our data on *ankyrin* suggest that per nucleotide site variability at *ankyrin* is reduced at least 20-fold compared to the mean value for genes in regions of normal recombination (Table 1). Chromosome 4 is achiasmate and shows little or no crossing over in *D. melanogaster* under normal conditions (Ashburner 1989, Chap. 11; Hawley *et al.* 1993), so that this lack of variability is in agreement with the general pattern of a correlation between the local recombination rate experienced by a gene and its level of sequence variability, observed in *D. melanogaster* (Aguadé and Langley 1994; Aquadro *et al.* 1994) and, increasingly, in other taxa including humans (Charlesworth and Charlesworth 1998; Nachman 2001). There is no evidence for an unusually low level of silent site divergence between *D. melanogaster* and *D. simulans* for *ankyrin*; in fact, the silent site divergence for *ankyrin* is about twice the mean value given by Moriyama and Powell (1996). This excludes the possibility that recombination is mutagenic, leading to reduced variation in regions on chromosome 4 because of lower mutational input, in agreement with previous studies (Begun and Aquadro 1992; Aguadé and Langley 1994). However, some recent results (W. Wang, K. Thornton, A. Berry and M. Long, personal communication) indicate that variability in another region of chromosome 4 is much higher than that at *ankyrin* and ci^D, possibly suggesting the action of balancing selection. Some recombination must be occurring on chromosome 4 if these observations are to be reconciled. As discussed below, we have direct evidence that this is the case.

Recombination and linkage disequilibrium: As described in results, the sequence data displayed in Tables 2 and 3 and Figures 3 and 4 show evidence for recombination events in both *D. melanogaster* and *D. simulans*, if we assume that the variants concerned represent unique mutations (Hudson and Kaplan 1985). This is consistent with other evidence from surveys of natural polymorphisms in Drosophila, which indicate the occurrence of recombination events even in regions where crossing over is very infrequent (Andolfatto and Nordborg 1998; Langley *et al.* 2000). Given the achiasmate behavior of chromosome 4 at meiosis (Hawley *et al.* 1993), these events most likely involve gene conversion rather than reciprocal exchange, although the DNA sequence data do not distinguish between the two. This is consistent with the interpretation of early recombination nodules in female meiosis as precursors of both types of recombination event and the fact that these are just as frequent in regions where crossing over is suppressed as in regions with normal frequencies of crossing over (Zickler and Kleckner 1999). This suggests that gene conversion may occur on the fourth chromosome and at the tips and bases of the major chromosomes.

A difficulty with this interpretation is that we find high levels of linkage disequilibrium between very distant sites, in contrast to what is found for the *su(s)* and $su(w^a)$ loci at the tip of the X chromosome, for which it has been suggested that gene conversion is the major factor involved in reducing linkage disequilibrium between sites within genes (Langley *et al.* 2000). Given that mean meiotic conversion tract lengths are thought to be of the order of 350 bp (Hilliker *et al.* 1994), sites at opposite ends of *ankyrin* should experience recombination due to gene conversion at maximal rates, yet inspection of Tables 2 and 3 shows several examples of complete or nearly complete linkage disequilibrium for pairs of sites >4 kb apart. For example, the squared correlation between sites 896 and 5366 in *D. melanogaster* is 0.89. It is possible that this simply reflects the lower level of variation at *ankyrin* compared to these loci [0.5×10^{-4} as opposed to an average of $>10^{-3}$ for *su(s)* and $su(w^a)$]. Since the effect of gene conversion on the expected magnitude of linkage disequilibrium depends on the product of $4N_e$ and the associated recombination parameter (Andolfatto and Nordborg 1998), the lower N_e associated with its lower level of variation may cause linkage disequilibrium to extend over a larger distance at *ankyrin* than at the X chromosomal loci. For example, even if the rate of recombination per base pair due to gene conversion were as high as 10^{-6} at *ankyrin*, an N_e of 50,000 (consistent with the 20-fold reduction in variability at *ankyrin*) would yield an expected value of the squared correlation between sites of $\sim 1/(1 + 4 \times 50{,}000 \times 10^{-6}) = 0.833$, consistent with the observed high values.

Causes of reduced variability at *ankyrin*: As discussed above, it is likely that some form of hitchhiking effect of selection on variability at linked sites has resulted in the observed pattern of reduced variation at *ankyrin*. One aim of this study was to attempt to discriminate between alternative versions of hitchhiking (see the Introduction). Table 1 gives no evidence for a significantly negative Tajima's *D* statistic (Tajima 1989a) for *ankyrin* (indeed, *D* is even positive in *D. melanogaster*), so that there is no evidence for the distorted nucleotide site frequency spectrum expected if there had been a sufficiently recent selective sweep to reduce variation to the observed extent (Braverman *et al.* 1995; Simonsen *et al.* 1995). In addition, the likelihood analyses described above are apparently inconsistent with a selective sweep as the explanation for the severely reduced variability that we observe (see Figures 5 and 6). There are, however, some difficulties in accepting this conclusion at face value. First, this study has uncovered evidence of some genetic exchange within the *ankyrin* locus (see above), so that it is appropriate to examine the possibility that there was a selective sweep with some recombination during the sweep. Fay and Wu (2000) pointed out that a selective sweep in regions with recombination can lead to a distinctive frequency spectrum of derived variants, in which a portion of the presweep variants

are sent to a band of low frequencies, because of their assocation with the allele that is eliminated from the population; another portion is sent to near fixation, as a result of preexisting variants recombining onto the haplotype that is destined for fixation. This process is, however, inconsistent with the intermediate-frequency-derived variants that we observe (for a detailed analysis, see Jensen 2000).

Furthermore, if there has been a sweep with recombination, this implies that we must assume that the time since the sweep (T_s) is nonzero. Suppose that we observe a sample and assume that it is the result of a sweep with zero recombination, as in our simulations. If there was in fact some recombination, some of the low frequency variants in the sample may be presweep variants that remained on the portion of the genealogy that recombined onto the selectively favorable allele; the remaining part of the sample represents the portion of the genealogy that was swept clean of its preexisting variation by the spread of the favorable allele (see Figure 2 of Fay and Wu 2000) and hence is equivalent to a truncated coalescent with no recombination. From the above argument, the "extra" presweep variants must be present at low frequencies in the sample and lead to an overestimate of the number of postsweep segregating sites, for a given T_s. It is, therefore, less likely that a null model of a selective sweep would be rejected by the likelihood method assuming a truncated coalescent, on the basis of a given neutral value of θ, making it more difficult to obtain the results shown in Figures 5 and 6. This means that the zero-recombination assumption is in fact conservative for our purposes.

The other problem with the tests for a selective sweep is our assumption of a constant postsweep population size in testing for the effects of a sweep. There is accumulating evidence that frequency spectra in non-African populations of both *D. simulans* and *D. melanogaster* may be distorted as a result of demographic effects, such as recent population bottlenecks associated with colonization events (Langley *et al.* 2000; Przeworski *et al.* 2001). While the values of statistics such as Tajima's *D* when averaged over loci tend to be close to neutral expectation (Przeworski *et al.* 2001), the stochastic nature of bottleneck effects means that different loci may be affected in different ways, which makes it difficult to conduct tests on a single locus that take account of bottlenecks. We note, however, that the absence of a strong reduction in variability at autosomal loci in non-African populations compared with African populations (Andolfatto 2001) suggests that any bottlenecks must have been only partial. With the approximately star-shaped phylogeny expected from a recent sweep, a partial bottleneck that occurred after the sweep, and that was then followed by almost instantaneous population expansion, could have the effect of causing some of the lineages that survived the bottleneck to be represented multiple times in the sample, so that any variants that they had acquired prior to the bottleneck would be represented at intermediate frequencies. This is qualitatively consistent with the data in Tables 2 and 3. But it is unclear on the basis of this hypothesis why such distortions toward intermediate frequencies are not detected more often at loci in regions of normal recombination (Przeworski *et al.* 2001). A study of a sample from an African population, which would be more likely to be close to equilibrium, would help to test the bottleneck hypothesis.

Alternatives to a selective sweep: Overall, our analysis suggests that the reduction in diversity on chromosome 4 in *D. melanogaster* and *D. simulans* is unlikely to have been caused by a selective sweep involving strong selection, unless the sweep was followed by a recent and partial population bottleneck. It is difficult to discriminate among other alternative hypotheses that might explain the reduced variability. Gillespie (1994, 1997) studied several models of temporally fluctuating selection coefficients, in terms of their effects on neutral variability at closely linked sites. His results suggest that the only process capable of producing the magnitude of reduction in variability that we have observed for chromosome 4 is the TIM model (Takahata *et al.* 1975). This involves temporal variation in selection coefficients without any component of balancing selection, such that a mutation eventually becomes fixed or lost over a timescale that is substantially shorter than the coalescent time but much longer than the time assumed in the catastrophic sweep model considered above. It causes only a moderate distortion of the allele frequency spectrum at linked neutral sites (Gillespie 1997), consistent with our observations.

The other possibilities are background selection (Charlesworth *et al.* 1993) and Hill-Robertson interference between weakly selected sites (McVean and Charlesworth 2000). Both of these can produce substantial reductions in variation in regions where recombination is greatly reduced. However, there is evidence that selection on silent sites is currently essentially absent in *D. melanogaster* (Akashi 1996; McVean and Vieira 2001), so it seems unlikely that weak Hill-Robertson effects can cause reduced variability in this species.

As far as background selection is concerned, it can produce reductions in variation without significantly skewing the sample frequency spectra at neutral sites, although it may produce a spectrum skewed in favor of rare variants if selection is very weak (Hudson and Kaplan 1994; Charlesworth *et al.* 1995). The detection of significant negative skews in regions of reduced recombination is therefore not conclusive evidence for selective sweeps, as is sometimes stated (Langley *et al.* 2000). As argued by Charlesworth (1996), the most important source of background selection for a small region of low recombination, like chromosome 4, is likely to be weak selection against transposable elements. The available population data on transposable elements in *D. melanogaster* are consistent with most ele-

ment families being maintained by an approximate balance between transposition and selection (Charlesworth *et al.* 1992a,b). Under such a balance, the mean selection coefficient against an element insertion must be equal to the transposition rate, which is typically of the order of 10^{-4} (Maside *et al.* 2000), so that the selection concerned is weak but much stronger than the reciprocal of the effective population size. This means that it is reasonable to treat the effect of transposable elements in the same way as deleterious mutations at equilibrium under mutation and selection. In the absence of recombination, this implies that neutral variability on chromosome 4 should be reduced below neutral expectation by a factor of $\exp(-n)$, where n is the mean number of elements per fourth chromosome (Charlesworth 1996). Data on *D. melanogaster* suggest a conservative estimate of at least 6.4 for n (Charlesworth *et al.* 1992b), so that the observed level of variability is >10-fold greater than predicted by this formulation. Element abundances in *D. simulans* are generally ~3-fold lower than in *D. melanogaster* (Biémont and Cizeron 1999), and preliminary data on our population sample indicate a correspondingly low copy number on chromosome 4 (M. Boulesteix and X. Maside, unpublished data). A copy number of around two to three per fourth chromosome would be, in fact, quite consistent with our observations. Since element abundances can change rather quickly over evolutionary time (Maside *et al.* 2000), it is possible that the high chromosome 4 copy number in *D. melanogaster* represents a recent situation and that neutral variability has not yet equilibrated to the effective population size corresponding to this copy number. This is supported by data suggesting lower mean transposable element copy numbers in African compared with non-African populations of *D. melanogaster* (Biémont *et al.* 2001).

History of the HB-related element insertion: The results described above show that the samples from both *D. melanogaster* and *D. simulans* are fixed for a copy of the transposable element HB (Brierly and Potter 1985; Harris *et al.* 1988; Henikoff 1992). In addition, we detected a probable transposable element insertion at a frequency of 9.1% in *D. simulans*. The difference between the two species in the location of HB implies that either a single insertion of HB into intron 2 occurred in the ancestral population, followed by a short-distance transposition that shifted its location in one or the other population after isolation, or that species-specific HB-like elements inserted in each lineage independently, after isolation. The second of these possibilities seems more likely, for the following reasons.

First, note that the pairwise divergences in Table 5 (standard *melanogaster* HB/*simulans* insertion, standard HB/*melanogaster* insertion, and *simulans*/*melanogaster*) are approximately equal and rather large (~13%). This suggests that the most recent common ancestor of the two *ankyrin* insertions is a relatively distant ancestor of all three elements, since otherwise we would expect at least one of the distances of HB/*simulans* or HB/*melanogaster* to be significantly less than the divergence between the species. If this is the case, it is likely that species-specific members of the HB family were responsible for the insertions; *i.e.*, that separate insertion events were involved. Also, the divergence between the *melanogaster* and *simulans* insertions is significantly greater than the divergence between the remaining homologous DNA in the region, which is inconsistent with a single insertion diverging at the same rate as the flanking DNA, if the mutation rate is uniform across the region.

More evidence for species-specific insertions is provided by the state of degeneration of the insertions. Table 6 shows that the *D. melanogaster* insertion has experienced eight deletions and an insertion with respect to the standard HB sequence, but that the *D. simulans* insertion has not experienced such events; three of these events involve the HB ORF. The *D. melanogaster* insertion also lacks any trace of the HB terminal inverted repeats. The *D. simulans* insertion, on the other hand, contains an entire ORF, with only a single nonsense mutation, and retains nearly intact TIRs. This may reflect much more recent fixation of the element at this location in *D. simulans*.

Overall, therefore, the data suggest that two independent insertions of HB occurred in the two species, into the *ankyrin* intron 2. Together with the relatively high frequency of another element in *D. simulans*, this is a very striking observation. Transposable element frequencies in *D. melanogaster* populations at individual chromosomal sites are almost universally low, except in proximal portions of the chromosome arms where recombination is greatly restricted (Charlesworth *et al.* 1992a,b; Biémont *et al.* 1997); however, even on chromosome 4, fixation of elements is uncommon (Charlesworth *et al.* 1992b). Element abundance is generally lower in *D. simulans* and some cases of apparent fixation (at the level of polytene chromosome bands, rather than nucleotide sites) have been reported (Biémont *et al.* 1997). The fixation of HB in the *ankyrin* intron is consistent with the pattern of an accumulation of elements in regions where crossing over is highly suppressed (Charlesworth *et al.* 1992a,b); the large size of this intron is due in part to the presence of this insertion and conforms to the general pattern of larger than average introns in regions of reduced crossing (Comerón and Kreitman 2000), although there is no evidence that this pattern usually involves transposable element insertions (Comerón and Kreitman 2000). It is interesting to note that the large introns of the Y chromosomal male fertility genes of *D. hydei* also contain large numbers of transposable elements, as well as satellite sequences (Hackstein and Hochstenbach 1995; Reugels *et al.* 2000). Population genetic mechanisms for the accumulation of repetitive DNA in regions of restricted crossing over have been discussed in the literature (Charlesworth *et al.* 1994). The fixation of HB

in *ankyrin* is thus consistent with this wider pattern of genome evolution.

We thank R. R. Hudson for help with the simulation methodology; C. Bergman, J. Comerón, J. Fay, C.-I Wu, and G. Wyckoff for stimulating discussions that improved this work; D. Guttman for expert advice and training; and M. Long and K. H. Jensen for generous support. P. Oefner, P. Underhill, T. Morton, and Transgenomic, Inc. provided invaluable help with DHLPC. J. Wall kindly analyzed our data using his program for estimating gene conversion rates. We also thank two anonymous reviewers for their comments. This work was supported by a National Institutes of Health Genetics and Regulation Training Grant predoctoral fellowship and a National Science Foundation doctoral dissertation improvement grant DEB-9701114 to M.A.J. B.C. is supported by the Royal Society.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

Aguadé, M., and C. H. Langley, 1994 Polymorphism and divergence in regions of low recombination in *Drosophila*, pp. 67–76 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, London.

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.

Aguadé, M., W. Meyers, A. D. Long and C. H. Langley, 1994 Single-strand conformation polymorphism analysis coupled with stratified DNA sequencing reveals reduced sequence variation in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome. Proc. Natl. Acad. Sci. USA **91:** 4658–4662.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino-acid substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297–1307.

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

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

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.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bandelt, H. J., P. Forster and A. Rohl, 1999 Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16:** 37–48.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B **355:** 1553–1562.

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.

Benson, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27:** 573–580.

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

Biémont, C., and G. Cizeron, 1999 Distribution of transposable elements in Drosophila species. Genetica **105:** 43–62.

Biémont, C., C. Vieira, C. Hoogland, G. Cizeron, C. Loevenbruck *et al.*, 1997 Maintenance of transposable element copy number in natural populations of *Drosophila melanogaster* and *D. simulans*. Genetica **100:** 161–166.

Biémont, C., N. Borie and C. Vieira, 2001 Transposable elements and genome evolution in *Drosophila melanogaster* and *D. simulans*. Abstract of the 17th European Drosophila Research Conference, September 2001, Edinburgh.

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.

Brierly, H. L., and S. S. Potter, 1985 Distinct characteristics of loop sequences of two Drosophila foldback transposable elements. Nucleic Acids Res. **13:** 485–500.

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

Charlesworth, B., A. Lapid and D. Canada, 1992a The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. I. Element frequencies and distribution. Genet. Res. **60:** 103–114.

Charlesworth, B., A. Lapid and D. Canada, 1992b The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II. Inferences on the nature of selection against elements. Genet. Res. **60:** 115–130.

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

Charlesworth, B., P. Sniegowski and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371:** 215–220.

Charlesworth, D., and B. Charlesworth, 1998 Sequence variation: looking for effects of genetic linkage. Curr. Biol. **8:** R658–R661.

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

Comerón, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila. Dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

Comerón, 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.

Dubreuil, R. R., and J.-Q. Yu, 1994 Ankyrin and beta-spectrin accumulate independently of alpha-spectrin in Drosophila. Proc. Natl. Acad. Sci. USA **91:** 10285–10289.

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

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Gillespie, J. H., 1994 Alternatives to the neutral theory, pp. 1–17 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

Gillespie, J. H., 2000 Genetic drift in an infinite population. The pseudohitchhiking model. Genetics **155:** 909–919.

Hackstein, J. H., and R. Hochstenbach, 1995 The elusive fertility genes of Drosophila: the ultimate haven for selfish genetic elements. Trends Genet. **11:** 195–200.

Harris, L. J., D. L. Baillie and A. M. Rose, 1988 Sequence identity between an inverted repeat family of transposable elements in Drosophila and Caenorhabditis. Nucleic Acids Res. **16:** 5991–5998.

Hawley, R. S., K. S. McKim and T. Arbel, 1993 Meiotic segregation in *Drosophila melanogaster* females, mechanisms and myths. Annu. Rev. Genet. **27:** 282–317.

Henikoff, S., 1992 Detection of Caenorhabditis transposon homologs in diverse organisms. New Biol. **4:** 382–388.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Hilliker, A. J., and A. Chovnick, 1981 Further observations on intragenic recombination in *Drosophila melanogaster*. Genet. Res. **38:** 281–296.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark *et*

al., 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics **137:** 1019–1026.

Hilton, H., R. M. Kliman and J. Hey, 1994 Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48:** 1900–1913.

Hochman, B., 1976 The fourth chromosome of *Drosophila melanogaster*, pp. 903–928 in *The Genetics and Biology of Drosophila*, Vol. 1b, edited by M. Ashburner and E. Novitski. Academic Press, New York.

Huber, C. G., P. J. Oefner, E. Preuss and G. K. Bonn, 1993 High-resolution liquid chromatography of DNA fragments on non-porous poly(styrene-divinylbenzene) particles. Nucleic Acids Res. **21:** 1061–1066.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

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, 1994 Gene trees with background selection, pp. 140–153 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

Jensen, M. A., 2000 Population genetics investigations in Drosophila and Saccharomyces. Ph.D Dissertation, University of Chicago, Chicago.

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

Keightley, P. D., and A. Eyre-Walker, 2000 Deleterious mutations and the evolution of sex. Science **290:** 331–333.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837–1852.

Maside, X., S. Assimacopoulos and B. Charlesworth, 2000 Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. Genet. Res. **75:** 275–284.

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 B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected sites on patterns of molecular evolution and variation. Genetics **155:** 929–944.

McVean, G. A. T., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245–257.

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

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–484.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Perlitz, M., and W. Stephan, 1997 The mean and variance of the number of segregating sites since the last hitchhiking event. J. Math. Biol. **36:** 1–23.

Pritchard, J. K., and S. W. Schaeffer, 1997 Polymorphism and divergence at a Drosophila pseudogene locus. Genetics **147:** 199–208.

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.

Reugels, A. M., R. Kurek, U. Lammermann and H. Bünemann, 2000 Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of *Drosophila hydei*. Genetics **154:** 759–769.

Rozas, M., and R. Rozas, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests on neutrality of DNA polymorphism data. Genetics **141:** 413–429.

Stephan, W., 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Mol. Biol Evol. **12:** 959–962.

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., 1951 A map of the fourth chromosome of *Drosophila melanogaster*, based on crossing over in triploid females. Proc. Natl. Acad. Sci. USA **37:** 405–407.

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.

Takahata, N., K. Ishii and H. Matsuda, 1975 Effect of temporal fluctuation of selection coefficient on gene frequency in a population. Proc. Natl. Acad. Sci. USA **72:** 4541–4545.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–80.

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

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **10:** 256–276.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–854.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Zickler, D., and N. Kleckner, 1999 Meiotic chromosomes: integrating structure and function. Annu. Rev. Genet. **33:** 603–754.

Communicating editor: W. Stephan