Molecular Variation at the *In(2L)t* Proximal Breakpoint Site in Natural Populations of *Drosophila melanogaster* and *D. simulans*

Peter Andolfatto and Martin Kreitman

Committee on Genetics, Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received May 8, 1999
Accepted for publication December 22, 1999

ABSTRACT

A previous study of nucleotide polymorphism in a Costa Rican population of *Drosophila melanogaster* found evidence for a nonneutral deficiency in the number of haplotypes near the proximal breakpoint of *In(2L)t*, a common inversion polymorphism in this species. Another striking feature of the data was a window of unusually high nucleotide diversity spanning the breakpoint site. To distinguish between selective and neutral demographic explanations for the observed patterns in the data, we sample alleles from three additional populations of *D. melanogaster* and one population of *D. simulans*. We find that the strength of associations among sites found at the breakpoint varies between populations of *D. melanogaster*. In *D. simulans*, analysis of the homologous region reveals unusually elevated levels of nucleotide polymorphism spanning the breakpoint site. As with American populations of *D. melanogaster*, our *D. simulans* sample shows a marked reduction in the number of haplotypes but not in nucleotide diversity. Haplotype tests reveal a significant deficiency in the number of haplotypes relative to the neutral expectation in the *D. simulans* sample and some populations of *D. melanogaster*. At the breakpoint site, the level of divergence between haplotype classes is comparable to interspecific divergence. The observation of interspecific polymorphisms that differentiate major haplotype classes in both species suggests that haplotype classes at this locus are considerably old. When considered in the context of other studies on patterns of variation within and between populations of *D. melanogaster* and *D. simulans*, our data appear more consistent with the operation of selection than with simple demographic explanations.

THE chromosomal rearrangement *In(2L)t* is one of four common polymorphic inversions with stable geographic frequency clines in natural populations of *Drosophila melanogaster*. While rare in temperate climates, it reaches frequencies of 40–60% in tropical populations of Australasia and Africa (Knibb 1982; Benassi *et al.* 1993). The existence of parallel latitudinal clines across different continents and hemispheres suggests that natural selection maintains at least some inversion polymorphisms in this species (Knibb 1982). The mechanism by which inversions become established in natural populations and the mode of selection operating on them are not well understood.

In a recent study of nucleotide variation spanning the proximal *In(2L)t* breakpoint, Andolfatto *et al.* (1999) found that the inversion has a recent origin relative to standard lineages. The authors also noted that standard chromosomes exhibit a several hundred-base pair window of elevated nucleotide polymorphism directly spanning the *In(2L)t* breakpoint site. Interestingly, most of this nucleotide variation in this window is distributed between, and not within, two deeply diverged standard haplotype classes. A much larger (~2 kb) DNA interval, which includes this window of elevated polymorphism, revealed little evidence for recombination despite the large number of intermediate frequency (*i.e.*, informative) polymorphic sites.

A departure from the neutral prediction for the number of haplotypes can arise as a result of selection or demographic shifts. In particular, a reduction in the number of haplotypes in a sample is expected under models of balancing selection or population subdivision (Strobeck 1987). Similar patterns are expected for partial selective sweeps (Maynard-Smith and Haigh 1974; Kaplan *et al.* 1989; Braverman *et al.* 1995; Hudson *et al.* 1997) or traffic models (Kirby and Stephan 1996). Under a neutral equilibrium model with recombination, the distribution of the expected number of haplotypes in a population sample can be determined by coalescent simulation (Hudson 1990; Fu 1996). Andolfatto *et al.* (1999) introduce a test of the neutral equilibrium model with recombination (based on Strobeck 1987) to detect subregions of a data set with unusual haplotype structure. A large window of polymorphisms spanning the *In(2L)t* breakpoint was shown to have fewer haplotypes than expected under the neutral model. This pattern is apparent whether *In(2L)t* chromosomes are included in the analysis or not.

Similar haplotype deficiencies have recently been reported at *Sod*, *vermilion*, *Fbp2*, and *Su(H)* in *D. melanogaster* (Hudson *et al.* 1994; Begun and Aquadro 1995;

Corresponding author: Peter Andolfatto, Institute of Cell, Animal and Population Biology, Ashworth Labs, Kings Bldgs., University of Edinburgh, Edinburgh, EH9 3JT Scotland, United Kingdom.
E-mail: peter.andolfatto@ed.ac.uk

Benassi *et al.* 1999; Depaulis *et al.* 1999) as well as at the *Pgd*, *runt*, *G6pd*, and *vermilion* loci in *D. simulans* (Begun and Aquadro 1994; Hamblin and Veuille 1999; Labate *et al.* 1999). One explanation is that the recent expansion of African populations of *D. melanogaster* and *D. simulans* to other parts of the world (David and Capy 1988; Lachaise *et al.* 1988) resulted in the deficiency in haplotype and nucleotide diversity observed in non-African populations (Hale and Singh 1991; Begun and Aquadro 1993, 1994, 1995). However, a characteristic feature of demographic shifts is that their signature is expected over the whole genome rather than localized to any particular locus. Interestingly, polymorphic sites further away from *Sod*, *Su(H)*, and the *In(2L)t* breakpoint in *D. melanogaster* reveal more recombination (Bénassi *et al.* 1993; Hudson *et al.* 1997; Andolfatto *et al.* 1999), consistent with expectations under certain selection models (Kaplan *et al.* 1989; Braverman *et al.* 1995; Kirby and Stephan 1996; Hudson *et al.* 1997).

Distinguishing between selective and demographic explanations for patterns of nucleotide diversity at a particular locus on the basis of data from a single sample is difficult. Here, we expand the polymorphism data set of Andolfatto *et al.* (1999) to include samples from three additional geographically diverse populations of *D. melanogaster* and one population of *D. simulans.* Our study focuses on a 1-kb region spanning the proximal breakpoint of *In(2L)t* that exhibits both a window of elevated polymorphism and strong linkage disequilibrium among sites in the Costa Rican sample. Data on the geographic distribution of variation at this locus in *D. melanogaster* and comparisons with *D. simulans* may help us distinguish among evolutionary scenarios. If the strong linkage disequilibrium observed in this population is due to ancient balancing or epistatic selection, we may expect to see similar haplotype structure in all *D. melanogaster* populations and, potentially, in *D. simulans.* We may also expect to see a number of *trans*-specific polymorphisms. Alternatively, if the pattern in Costa Rica is the result of a recent contraction in population size as suggested by data from the X chromosome (Begun and Aquadro 1993), we expect to observe reduced nucleotide variation in non-African samples relative to African samples. Finally, if the unusual haplotype structure in the Costa Rican *D. melanogaster* sample is due to the recent increase in *In(2L)t*'s frequency (Andolfatto *et al.* 1999), we would not expect a similar pattern in a population sample of *D. simulans* (which lacks *In(2L)t*).

MATERIALS AND METHODS

Population samples and sequencing: The isolation of the *In(2L)t* proximal breakpoint (34A8–9 on the cytological map) and the collection of polymorphism data from a San Jose, Costa Rica *D. melanogaster* population sample are described by Andolfatto *et al.* (1999). Three additional population samples are from Florida City, Florida, Yeppoon, Australia, and Zimbabawe, Africa and were chosen primarily because they are geographically diverse. Only standard alleles are sampled in this study; *In(2L)t* alleles sampled from all four populations are described in Andolfatto *et al.* (1999).

Genomic DNA was prepared from wild-caught females from the Florida City population. Individuals were karyotyped by PCR with the use of standard and *In(2L)t*-specific primer pairs (Andolfatto *et al.* 1999). For Yeppoon and Zimbabawe (which included individuals from both Harare and Sengwa), we chose one *In(2L)t* heterozygote male per isofemale line (kindly provided by C.-I Wu). A 1-kb segment spanning the inversion breakpoint site was PCR amplified from *D. melanogaster In(2L)t* heterozygotes, using standard-specific primers (see Figure 1). To obtain alleles from a *D. simulans* population (Arena Farms, Maryland), the following cross was carried out: Multiple males from each isofemale line were crossed to virgin female *In(2L)t* homozygotes of *D. melanogaster.* The resulting hybrid progeny (all female) were heterozygous for *In(2L)t.* This allowed the recovery of individual *D. simulans* alleles by PCR with standard arrangement-specific primers (one individual per isofemale line).

Polyethylene glycol (PEG)-precipitated templates were directly sequenced on both strands using a dRhodamine Terminator Cycle sequencing kit (Applied Biosystems, Foster City, CA) and run on an ABI377XL Automated Sequencer. Sequences were analyzed with ABI Sequence Analysis v3.0 software; contigs were managed with Sequencher v3.0 software. Sequences collected in this study have been deposited into GenBank under accession nos. AF217926–AF217949. Intraspecific alignments have been deposited into the EMBL database (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/) under accession nos. DS41064–DS41065.

Polymorphism analyses: Although we report all segregating polymorphisms (Figures 2 and 3), we have restricted our analyses to two-state single-nucleotide polymorphisms and insertion-deletions within each population (see Table 1). The neutral mutation parameter $\theta = 4N_e\mu$, where N_e is the effective population size of the species and μ is the neutral mutation rate, is estimated from both π, the average pairwise difference per base pair (Tajima 1983), and S, the number of polymorphic sites in the sample (θ_w; Watterson 1975). Certain analyses of polymorphism and divergence were performed with DnaSP v3.0 software (Rozas and Rozas 1999).

The population recombination rate, $C = 4N_e r$, where r is the recombination rate per base pair per generation, is estimated in three ways. A lower bound for C is based on the minimum number of inferred recombination events in the history of a sample (R_M, in Hudson and Kaplan 1985). C_{min} is defined as the highest value of C such that <2.5% of simulated data sets have R_M or more inferred recombination events. This estimate is not strictly conservative when used in our haplotype test (see below) but is useful since it represents a lower bound for the recombination rate (Hudson and Kaplan 1985; Wall 1999). To estimate C_{min}, coalescent simulations were carried out for a neutral panmictic population conditional on the sample size, n, the number of segregating sites, S, and the population recombination rate, C (Hudson 1993). A second estimate, C_{hud}, is an estimate of the expected population recombination rate and is obtained from polymorphism data by the method of Hudson (1987). C_{hud} is not employed in statistical tests. A third estimate of C is $C_{lab} = 4N_e\rho(1 - 2q(1 - q))$, where ρ is the estimate of rates of crossing over per base pair per generation for the cytological band 34A based on laboratory crosses (Comeron *et al.* 1999) and q is the estimated frequency of *In(2L)t*; N_e is taken to be 10^6 (Kreitman 1983). Laboratory estimates of recombination (ρ) based on

the exchange of distant flanking markers interpolated to the intragenic scale (*i.e.*, several kilobases) are likely to be underestimates of the true rate of exchange (r), since they ignore the added contribution of gene conversion (Andolfatto and Nordborg 1998). C_{lab} is not conservative in statistical tests but has the advantage of being independent of the sampled data and is our best *a priori* guess at the true population recombination rate in the chromosomal region studied.

Statistical tests of neutral equilibrium and panmictic population models: Tajima's D statistic (Tajima 1989) is used to characterize the skew in the frequency distribution of segregating mutations in our samples. To test for geographic differentiation between population samples of standard chromosomes, we use permutation tests described by Hudson *et al.* (1992a). Differentiation between populations for haplotype frequencies is measured by the statistic χ^2 (Nei 1987, p. 110). The statistic K^* (Hudson *et al.* 1992a) is based on the frequencies of individual segregating sites in two (or more) populations. These two statistics were the most powerful under all parameters considered in Hudson *et al.* (1992a). For each test, 100,000 permutations of the data were carried out. We report the one-tailed probability that the two samples were drawn from a single panmictic population. A program to perform these tests was kindly provided by R. Hudson. For comparisons to earlier studies, we also report F_{st} (Hudson *et al.* 1992b) although significance levels are not assessed for this statistic.

We use the haplotype test of Andolfatto *et al.* (1999) to detect deviations from the neutral model in the number of observed haplotypes in our population samples. Given a polymorphism data set with n chromosomes and S segregating sites, we define S_k to be the largest number of consecutive segregating sites that contain only k different haplotypes ($1 < k < n$). An empirical distribution of S_k is determined from 10,000 simulations using an infinite-sites, panmictic coalescent model conditional on n, S, and C (Hudson 1993). We then calculate the proportion, p_k, of simulated data sets that contain at least one stretch of S_k consecutive segregating sites having k or fewer haplotypes. This is equivalent to calculating the proportion of simulated data sets that have S_k greater than or equal to S_k observed in the data. Since choosing any particular value of k is arbitrary, we correct for the implicit multiple tests involved. This corrected P value is determined from further coalescent simulations that compare the actual smallest p_k value with simulated smallest p_k values. For *D. melanogaster* populations, haplotype tests were performed on constructed random samples based on the population's frequency of *In(2L)t* (Andolfatto *et al.* 1999). The presence of an inversion in our *D. melanogaster* samples makes estimators of C difficult to interpret. For this reason, the robustness of our results are tested over a wide range of values for C.

RESULTS

Polymorphism and recombination among *D. melanogaster* samples: The sampled region spans the proximal *In(2L)t* breakpoint (Figure 1) and includes 840 bp from region C (within the inverted region) and 160 bp from region D (outside the inverted region). The inversion breakpoint is located between positions 840 and 934 where all sampled *In(2L)t* chromosomes are fixed for a 94-bp deletion (Andolfatto *et al.* 1999). Nucleotide and insertion-deletion variation found in four population samples of *D. melanogaster* standard chromosomes is summarized in Figure 2. Our estimates of π, θ, and C/θ (Table 1) include all biallelic variation in a given

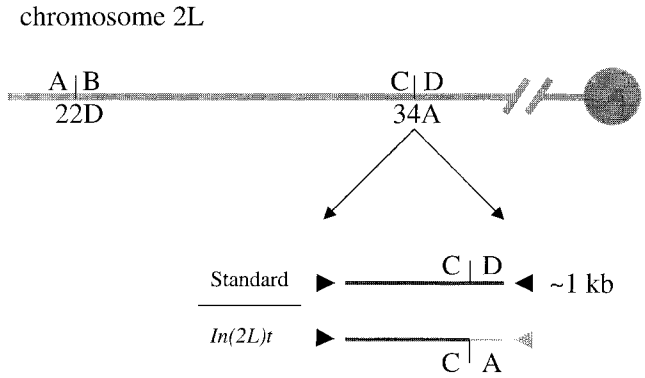

Figure 1.—PCR sampling strategy of a 1-kb region of the proximal breakpoint region from *In(2L)t* heterozygotes. The cytological position of this region (C/D) is ~34A8–9 on chromosome 2L. Standard-specific primer pairs were used to PCR amplify standard alleles from individual inversion heterozygotes.

population sample (see materials and methods). Two polymorphisms detected in the Florida City sample (polymorphic sites 77 and 83) as well as one in the Yeppoon sample (polymorphic site 5) were previously interpreted as fixed differences between standard and *In(2L)t* lineages (Andolfatto *et al.* 1999). Thus, since they are fixed in the *In(2L)t* class, these polymorphic sites are at intermediate frequencies in all populations. Polymorphic site 77 is a *trans*-specific polymorphism (see below).

Two features of the data differ between population samples of *D. melanogaster* (Figure 2). First, the Costa Rican and Miami samples appear to have stronger linkage disequilibrium among sites than Yeppoon and Zimbabwe samples. However, in contrast to previous studies for X-linked loci in various populations of *D. melanogaster* (Begun and Aquadro 1993, 1994, 1995), there is no evidence for an African / non-African difference in levels of nucleotide diversity (Table 1). Second, the frequency spectrum of polymorphisms is sharply skewed toward intermediate frequency mutations in the Costa Rican sample (Table 1). This skew is much less dramatic for the Florida City and Yeppoon samples and is in the direction of an excess of rare polymorphisms in the Zimbabwe sample. Tajima's D is significantly positively skewed for the Costa Rican sample for all $C \geq 0$ [constructed random samples (CRS), Table 1]. However, the interpretation of this P value is difficult since this sampled region was preselected because it appeared unusual in this population. Tajima's D was not significantly skewed for CRS of any other population when $0 \leq C \leq C_{lab}$.

Estimates of the population recombination rate also differ among samples. Under neutral equilibrium assumptions, the ratio C_{hud}/θ_w (Table 1) is an estimate of the expected number of recombination events per mutation in the sample. An independent measure of this quantity based on laboratory estimates of the recom-

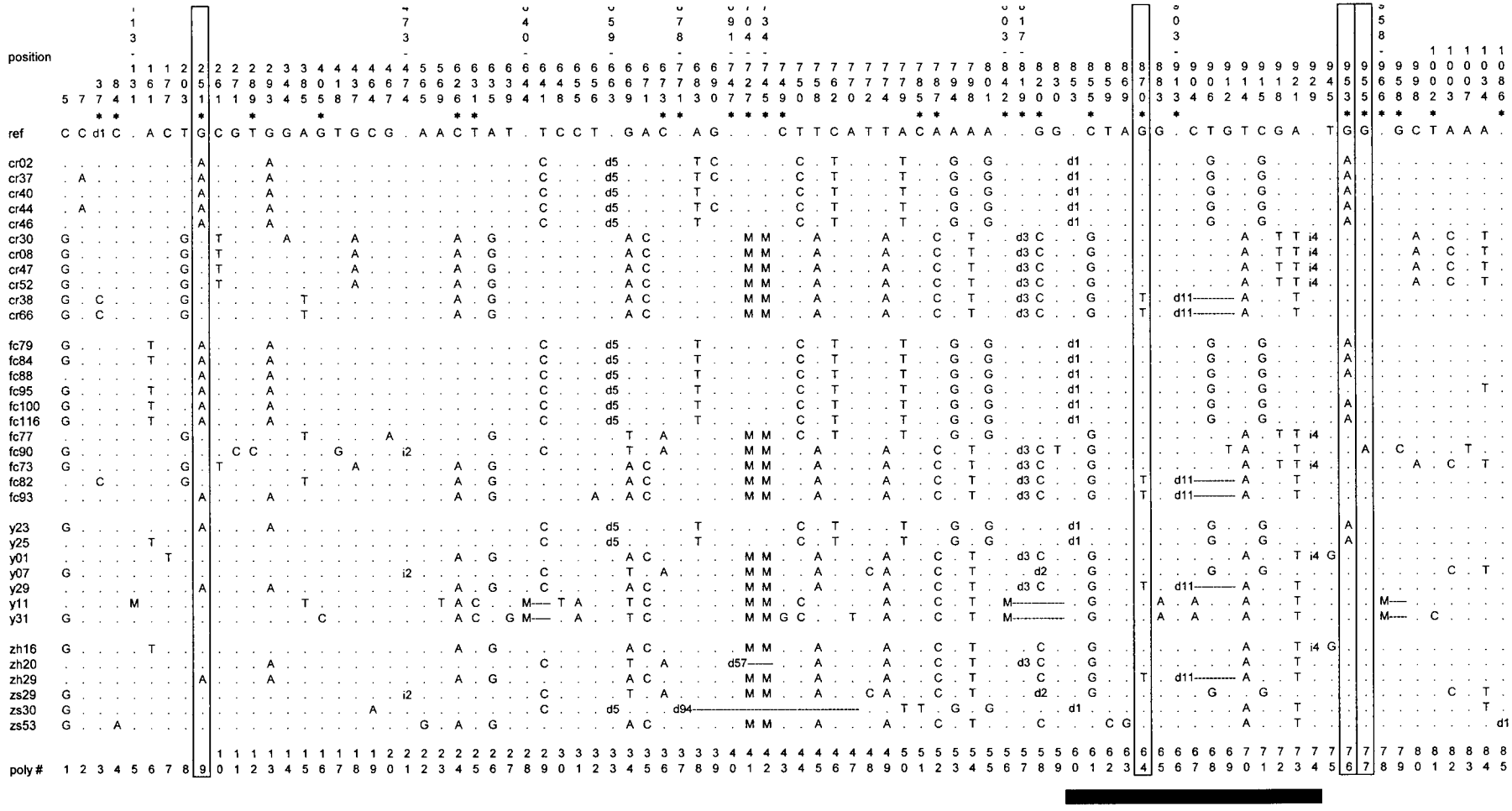

Figure 2.—Summary of polymorphic variation found in four population samples of *D. melanogaster* (cr, Costa Rica; fc, Florida City; y, Yeppoon; zh, Harare, Zimbabwe; zs, Sengwa, Zimbabwe). All mutations have been polarized using *D. simulans* as an outgroup; the reference sequence represents the ancestral state where possible (those sites with ambiguous polarity are indicated with an asterisk). M denotes complex mutations; i and d are insertions and deletions of the length indicated, respectively. Sites with more than two states or that overlap with deletions were excluded from analyses. The black bar indicates the position of the *In(2L)t* breakpoint, where inverted chromosomes are fixed for a 94-bp deletion. Polymorphism number 1 (nucleotide position 5) corresponds to polymorphism 76 (nucleotide position 2012) in Figure 2 of Andolfatto *et al.* (1999). Polymorphic sites 5, 77, and 83 were previously interpreted as fixed differences between karyotypes in the study of Andolfatto *et al.* (1999). Polymorphic site numbers 9, 64, 76, and 77 (boxed columns) are shared with *D. simulans* (Figure 3).

TABLE 1

Summary information for *D. melanogaster* and *D. simulans* population samples

Population	n^a	No. of sites[b]	S^c	θ_w	π	Tajima's D	$R_M{}^d$	C_{hud}/θ_w
D. melanogaster								
Standard alleles								
Costa Rica	11	987	43	0.015	0.022	2.3	0	0.13
Florida City	11	985	51	0.018	0.021	0.7	6	0.28
Yeppoon	7	957	48	0.021	0.021	0.1	4	2.29
Zimbabwe	6	927	36	0.017	0.015	−0.7	3	9.00
CRS[e]								
Costa Rica	14	742	33	0.014	0.021	2.0	0	0.14
Florida City	13	742	37	0.016	0.016	0.2	4	0.19
Yeppoon	9	729	34	0.017	0.018	0.2	2	0.52
Zimbabwe	7	688	21	0.013	0.012	−0.4	1	0.23
D. simulans								
Maryland	11	881	79	0.031	0.035	0.7	3	0.45

[a] Number of sampled chromosomes.
[b] Number of sites excluding alignment gaps in the sample.
[c] All biallelic segregating mutations including insertion-deletion polymorphism.
[d] Minimum number of recombination events in the sample (Hudson and Kaplan 1985).
[e] Constructed random samples based on *In(2L)t* frequencies estimated in each population (Costa Rica 20.8%; Florida City 25.0%; Yeppoon 22.9%; Zimbabwe 58.2%; Andolfatto *et al.* 1999).

bination rate in this chromosomal region ($\rho = 1.47 \times 10^{-8}$ per base pair per generation; Comeron *et al.* 1999) and the neutral mutation rate ($\mu \sim 1.6$–3.0×10^{-9} per site per generation, assuming 10 generations per year; Harada *et al.* 1993; Li 1997) yields a ratio (ρ/μ) of ~4.5–9.0. If we assume an *In(2L)t* frequency of 50% and no recombination in inversion heterozygotes, this range becomes ~2.3–4.5. While the Zimbabwe and Yeppoon samples of standard chromosomes are roughly in agreement with these estimated ranges, Costa Rica and Florida City samples yield a C_{hud}/θ_w ratio >10-fold smaller than expected (Table 1). Since C_{hud} is a summary of the amount of linkage disequilibrium in a sample (Hudson 1987) and population samples have similar estimates of θ_w, lower than expected C_{hud}/θ_w reflect stronger linkage disequilibrium in the American samples relative to Yeppoon and Zimbabwe samples. The C_{hud}/θ_w ratios are uniformly low in CRS (Table 1), likely due to the linkage disequilibrium introduced by *In(2L)t* chromosomes.

Geographic differentiation between population samples of standard chromosomes: Pairwise F_{st} estimates (Table 2, top right) suggest that differentiation between populations is low. Permutation tests described by Hudson *et al.* (1992a) were conducted on standard chromosome samples to test a panmictic population model (Table 2). In pairwise comparisons, the χ^2 statistic reveals haplotype differentiation between Costa Rica and all other populations. No other significant differentiation between haplotypes (χ^2) is detected in pairwise comparisons of populations. Two of six tests based on site frequencies (K^*) have P values near the 0.05 level (Zimbabwe/Costa Rica and Zimbabwe/Florida City). Caution should be exercised in interpreting the P values in Table 2 since they are not corrected for multiple tests. However, these tests do suggest haplotype differentiation between standard chromosomes of Costa Rican and other populations.

Polymorphism patterns in a *D. simulans* population sample: Figure 3 summarizes polymorphism data for the *In(2L)t* proximal breakpoint homologue in a North American population of *D. simulans*. Estimates of π, θ, and C/θ are given in Table 1. In agreement with previous data from nucleotide variation in *D. melanogaster* and *D. simulans* (reviewed in Moriyama and Powell 1996), estimates of θ from π and S are approximately twofold larger in *D. simulans* than in *D. melanogaster* (Figure 2). Unexpectedly, strong associations among polymorphic sites are observed in the *D. simulans* data set, similar to those seen in North American *D. melanogaster* samples. The *D. simulans* data set contains only three detected recombination events (by a four gamete test; Hudson and Kaplan 1985), despite a large number of informative sites. Tajima's D (Table 1) is positive but not significant under conservative estimates of recombination. Levels of nucleotide diversity in *D. simulans* show marked variation across the sequenced region (Figure 4a). As with *D. melanogaster*, a window spanning the *In(2L)t* breakpoint site homologue in *D. simulans* exhibits much higher levels of nucleotide diversity than average (~1.6 and 3.3% for silent sites in *D. melanogaster* and *D. simulans*, respectively; Moriyama and Powell 1996).

Divergence between haplotypes matches interspecific divergence: There is evidence that the major haplotype classes in both species are old relative to divergence between species. Figure 4, b and c shows levels of diver-

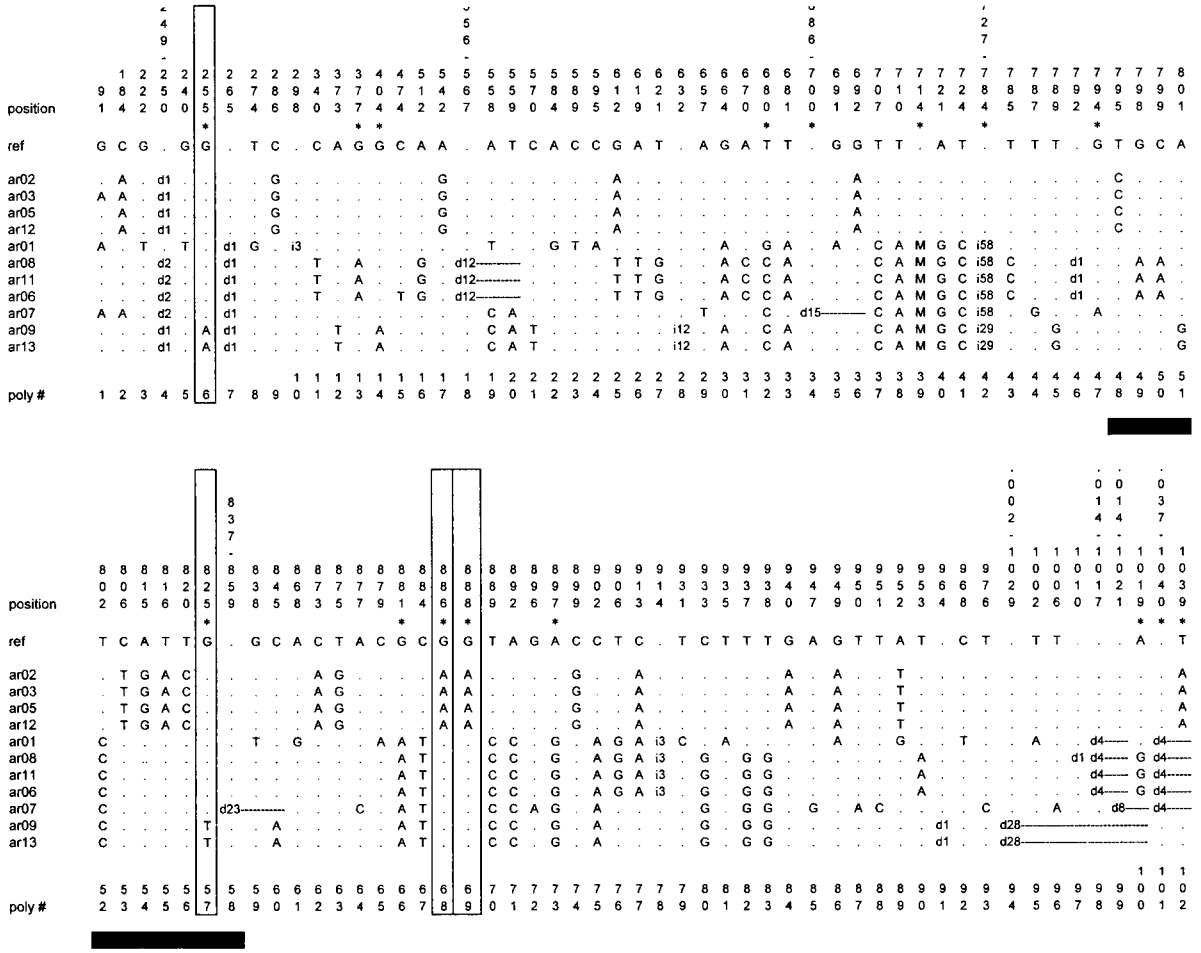

Figure 3.—Summary of polymorphic variation found for the *In(2L)t* proximal breakpoint site homologue in a Maryland population of *D. simulans*. All mutations have been polarized using *D. melanogaster* as an outgroup; the reference sequence represents the ancestral state where possible (those sites with ambiguous polarity are indicated with an asterisk). The black bar represents the approximate position of the *In(2L)t* breakpoint in *D. melanogaster*. Sites with more than two states or that overlap with deletions were excluded from analyses. Polymorphic sites 6, 57, 68, and 69 (boxed columns) are shared with *D. melanogaster* (Figure 2).

gence between two "haplotype classes" of the Costa Rican sample of *D. melanogaster* and our sample of *D. simulans*. Uncertainty in the alignment between the two species, especially near the breakpoint site, makes a quantitative assessment of divergence (and shared polymorphisms) difficult. However, a tentative alignment between *D. melanogaster* and *D. simulans* sequences reveals (qualitatively) that the breakpoint region has elevated interspecific divergence (Figure 4, b and c). Strikingly, average pairwise divergence between the two haplotype classes in both species matches or exceeds the estimated divergence between species near the *In(2L)t* breakpoint site. These observations can only be considered qualitative both because the alignment between *D. melanogaster* and *D. simulans* is poor and because the assignment of alleles to haplotype classes is arbitrary.

Shared polymorphisms: Shaded columns in Figures 2 and 3 show the positions of four *trans*-specific polymorphisms (sites 9, 64, 76, and 77 in Figure 2; sites 6, 57, 68, and 69 in Figure 3). All are at intermediate frequency (*i.e.*, sampled more than once) in *D. simulans* and CRS of *D. melanogaster*. In *D. melanogaster*, two of the four polymorphisms (9 and 76, Figure 2) differentiate two major standard haplotype classes in the Costa Rican population (AGAG and GKGG). Although sampled only once in the Florida City sample, polymorphism 77 is fixed in *In(2L)t* chromosomes and forms an inversion-specific haplotype (AGAA; Andolfatto *et al.* 1999). In *D. simulans*, two of the four polymorphisms (68 and 69, Figure 3) define two major haplotype classes in the *D. simulans* sample (GGAA and RKGG). The haplotypes formed by these polymorphisms are to some extent *trans*-specific. The ATGG haplotype, sampled twice in the *D. simulans* sample (Figure 3), is also found in the Florida City, Yeppoon, and Zimbabwe samples of *D. melanogaster*. The GGGG haplotype is found at intermediate frequency in the *D. simulans* sample and all *D. melanogaster* samples. One of these haplotypes (*i.e.*, ATGG or GGGG) may be ancestral.

Analysis of haplotype structure: We use the haplotype

TABLE 2

Geographic differentiation among standard chromosomes in *D. melanogaster*

	Costa Rica	Florida City	Yeppoon	Zimbabwe
Costa Rica	—	0.008	−0.007	0.045
Florida City	0.0388 0.1484	—	0.022	0.090
Yeppoon	0.0160 0.1165	0.1438 0.1830	—	−0.071
Zimbabwe	0.0145 0.0492	0.1192 0.0499	—[a] 0.8595	—

Estimates of F_{st} appear in the top right. One-tailed probabilities are reported for χ^2 (top) and K^* (bottom) in the bottom left. A total of 10,000 permutations were carried out following Hudson *et al.* (1992a). Only biallelic polymorphisms were included in the analyses.

[a] The χ^2 test was not performed in this case because each sampled chromosome represented a unique haplotype.

test of Andolfatto *et al.* (1999) to determine whether haplotype structure in our *D. simulans* sample is unusual under a neutral equilibrium model. The data depart from the neutral model when $C = C_{min}$ ($P = 0.029$, see Table 3). P values were significant for all simulations with $C \geq 0$ ($P < 0.035$) and P decreased monotonically for $C > C_{min}$. For comparison, we constructed CRS for each *D. melanogaster* population based on its estimated *In(2L)t* frequencies; 20.8% for Costa Rica, 25.0% for Florida City, 22.9% for Yeppoon, and 58.2% for Zimbabwe (Andolfatto *et al.* 1999). The neutral model was rejected (Table 3) for both the Costa Rican and Florida City populations when $C = C_{min}$; $P = 0.012$ and $P = 0.043$, respectively. For both populations, the neutral model was rejected in all simulations when $C \geq 0$ (maximum probabilities: Costa Rica, $P < 0.015$; Florida City, $P < 0.046$). Similar tests on CRS for Yeppoon and Zimbabwe (Table 3) do not reject the neutral model when $C = C_{min}$. When we condition simulations on $C = C_{lab}$, all probabilities become much lower; the null model is rejected for the Yeppoon sample assuming $C = C_{lab}$ ($P = 0.0175$).

DISCUSSION

Elevated polymorphism at the *In(2L)t* breakpoint site among standard chromosomes: Sliding-window analyses of polymorphism in the *D. melanogaster* (Costa Rica) and *D. simulans* samples reveal unusually high levels of nucleotide diversity near the *In(2L)t* breakpoint site (Figure 4a). This pattern is reminiscent of the predicted signature of balancing selection (Hudson and Kaplan 1988; Kreitman and Hudson 1991). A plausible alternative to balancing selection is simply that different DNA regions differ either in mutation rates or in levels of selective constraint. Variation in mutation rates across the sequence is unlikely given the observation that segregating mutations on *In(2L)t* chromosomes (not shown here) do not cluster near the inversion breakpoint (Andolfatto *et al.* 1999).

An argument in favor of heterogeneity in selective constraint across the region is that species polymorphism and divergence appear to be coupled (Figure 4). Indeed, the region immediately spanning the breakpoint appears to have elevated divergence as well as elevated levels of polymorphism (but seemingly lower divergence than that observed between major haplotype classes within each species). Comparisons of polymorphism and divergence for the *In(2L)t* proximal breakpoint with other loci (*i.e.*, the HKA test of Hudson *et al.* 1987) suggest that heterogeneity in constraint is a sufficient explanation for diversity levels in both species (results not shown). In addition, the analysis of a larger region surrounding the breakpoint (8.4 kbp) revealed at least two candidate exons (Andolfatto *et al.* 1999). The orientation of these putative exons is consistent with the presence of an additional exon or regulatory region in the 5′ region of the 1-kb region investigated here.

While heterogeneity in levels of constraint may be sufficient to explain levels of nucleotide diversity in the *In(2L)t* breakpoint region, it cannot explain the unusual distribution of this variation among haplotypes. This feature of the data is difficult to reconcile with a neutral equilibrium model given the rate of recombination expected in this chromosomal region (*i.e.*, C_{lab}; see Table 3). In addition, the several *trans*-specific polymorphisms in this region tend to fall on the deepest branches of genealogies in samples from both species. This observation can be taken as evidence that some component of this elevated window of nucleotide diversity surrounding the breakpoint is due to the long persistence of polymorphisms rather than simply a higher substitution rate (as suggested by the elevated divergence).

Geographic patterns and haplotype structure in *D. melanogaster*: We detect significant geographic differen-

tiation of standard haplotypes between populations of *D. melanogaster* despite low values of F_{st} (Table 2). The simplest explanation for the marked reduction in haplotype diversity in American samples is the recent expansion of African populations into the Americas (David and Capy 1988; Lachaise *et al.* 1988) resulting in founder effects or the recent admixture of previously subdivided populations. Several lines of evidence conflict with these simple demographic explanations for the *D. melanogaster* data. First, recent data from microsatellite loci (Irvin *et al.* 1998) offer no evidence for recent bottlenecks in non-African populations of *D. melanogaster*. Second, the four populations sampled do not differ from each other in levels of nucleotide diversity at the *In(2L)t* breakpoint (Table 2). When the presence of *In(2L)t* in each population is accounted for (CRS, Table 1), diversity is actually lowest in the African population. Several other autosomal loci show similar patterns (*Gld*, Hamblin and Aquadro 1997; *Acp26A*, Tsaur *et al.* 1998; *Adh*, S. C. Tsaur, unpublished results; *Tra*, R. Kulathinal, personal communication; but see Aguadé 1998, 1999). Thus, the apparent split between African and non-African populations reported for X-linked genes (Begun and Aquadro 1993–1995) does not seem to generalize to the autosomes. Third, the three non-African populations of *D. melanogaster* sampled for the *In(2L)t* breakpoint show distinctly different patterns of polymorphism (see Table 2). For example, Florida City and Yeppoon samples, in contrast to Costa Rica, reveal considerably more evidence for recombination and have less extreme values of Tajima's *D*. Finally, while both reduced haplotype diversity and a skew toward high frequency variants are observed in the Costa Rica data set, a survey of a larger region surrounding the *In(2L)t* breakpoint revealed lower frequency polymorphisms and more evidence for recombination (Andolfatto *et al.* 1999).

Thus, whether considering more populations, loci, or sites, one is led to some feature of the data that makes simple demographic explanations for the *D. melanogaster* data unlikely. It could be argued that, under demographic models with intermediate levels of recombination, a large variance in patterns of polymorphism is expected after a bottleneck (R. Hudson, personal communication). All demographic models, however, predict a reduction in the expected genome-wide level of polymorphism in bottlenecked populations. This pattern is not generally observed in the available autosomal data for *D. melanogaster*.

Comparing patterns in *D. melanogaster* and *D. simulans*: The observation of fewer haplotypes than ex-

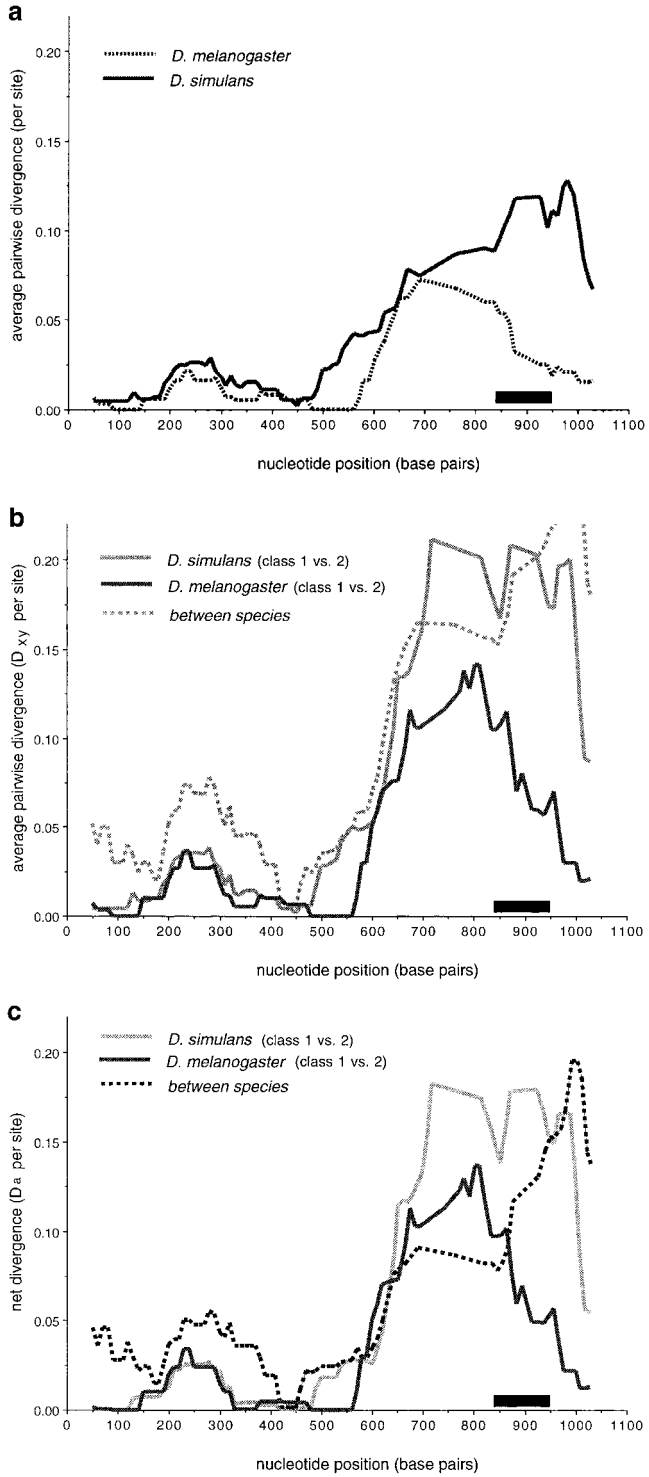

Figure 4.—(a) Sliding window of nucleotide diversity (π) across the sequenced region in *D. melanogaster* (dotted line) and *D. simulans* (solid line). All windows have an equal number of sites; the window size (excluding gaps) is 100 bp and the increment 10%. The approximate position of the *In(2L)t* breakpoint deletion is indicated by the black bar (positions 840–934, Figure 2). Nucleotide positions do not correspond exactly to those in Figures 2 and 3. (b and c) Average pairwise divergence between major haplotype classes of *D. melanogaster* and *D. simulans vs.* interspecific divergence (dotted line). Haplotype classes were arbitrarily defined: *D. melanogaster* class 1 alleles are cr30, cr08, cr47, cr52, cr38, and cr66 (Figure 2); *D. simulans* class 1 alleles are ar02, ar03, ar05, and ar12 (Figure 3). D_{xy} and D_a are average pairwise divergence and net divergence, respectively (*cf.* Nei 1987).

TABLE 3

Summary of haplotype tests on *D. melanogaster* and *D. simulans* populations

Population	k	S_k[a]	P[b] $C = 0$	$C = C_{min}$	$C = C_{lab}$
D. melanogaster					
Costa Rica	2	13			
(C. America)	3	21	0.0121	0.0121	<0.0001
	4	23			
	5	24			
	6	28			
Florida City	2	11	0.0451	0.0434	0.0020
(N. America)	3	14			
	4	18			
	5	27			
	6	33			
Yeppoon	2	6			
(E. Australia)	3	7			
	4	9			
	5	31	0.2301	0.2020	0.0175
	6	32			
Zimbabwe	2	4			
(C. Africa)	3	7			
	4	16	0.4290	0.4270	0.1730
	5	21			
D. simulans	2	6			
(N. America)	3	10			
	4	15			
	5	50	0.0324	0.0286	0.0004[c]
	6	78			
	7	79			

Haplotype tests for *D. melanogaster* were performed on constructed random samples (see Table 1). Boxed rows correspond to the window size (S_k) for which p_k is minimal.

[a] The longest number of consecutive polymorphic sites with k haplotypes.

[b] P is the one-tailed probability corrected for multiple tests and a *post hoc* choice of window size. C_{min} represents a lower bound for the recombination rate; C_{lab} is based on the laboratory estimates of the recombination rate (corrected for the expected suppression caused by *In(2L)t*) and assumes $N_e = 10^6$ (see materials and methods).

[c] We use the estimate of C_{lab} from *D. melanogaster*.

pected for the *In(2L)t* proximal breakpoint homologue in *D. simulans* could again be taken as evidence for a recent range expansion of African populations (as for *D. melanogaster* above). For example, data for *vermilion* and *G6pd* loci show evidence for reduced haplotype diversity in some non-African populations of *D. simulans* relative to African populations (Hamblin and Veuille 1999). It has been repeatedly suggested that the data in *D. simulans* reflect recent population admixture (Hasson *et al.* 1998; Hamblin and Veuille 1999; Labate *et al.* 1999).

An increasing number of reports of unusual haplotype structure in population samples of both *D. melanogaster* and *D. simulans* point to simple explanations, such as those based on the demographic histories of these species. However, while *D. melanogaster* and *D. simulans* may have certain similarities in their demographic histories (*i.e.*, a possible recent expansion from Africa), it seems unlikely that these histories will be similar enough to produce identical patterns at any particular locus under investigation. While multiple loci reveal evidence for geographic differentiation in both *D. melanogaster* and *D. simulans* (*e.g.*, Hale and Singh 1991; Begun and Aquadro 1993–1995; Hamblin and Veuille 1999), the pattern of haplotype structure varies from locus to locus. For example, while some loci (*e.g.*, *Pgd*, this study) show a nonneutral deficiency of haplotypes in North American populations of *D. simulans*, others (*e.g.*, *vermilion*, *Gld*) do not (Begun and Aquadro 1995; Hamblin and Aquadro 1996). Also, in contrast to the pattern reported here, unusual patterns at these loci are not *trans*-specific. For example, the *vermilion* locus shows a reduction in nucleotide diversity and number of haplotypes in a North American population of *D. melanogaster* but not *D. simulans* (Begun and Aquadro 1995).

Selection-based explanations for the unusual linkage disequilibrium patterns observed at the *In(2L)t* breakpoint site should also be entertained. The expansion of African populations of *D. melanogaster* and *D. simulans* into more temperate climates may have been accompanied by selection at many loci. In regions of intermediate recombination, this could lead to considerable heterogeneity in haplotype structure both among loci and among populations. This is a reasonable interpretation of the pattern seen in different populations of *D. melanogaster* at the *In(2L)t* breakpoint. It is also possible that the elevated nucleotide diversity, deficiency of haplotypes, and *trans*-specific polymorphisms that differentiate major haplotype classes are the result of long-standing epistatic interactions. Evidence for selective constraints and putative exons near the *In(2L)t* breakpoint (Andolfatto *et al.* 1999) suggests that this region is itself a potential target of selection. The relatively recent appearance of *In(2L)t* (Andolfatto *et al.* 1999) may seem to preclude its relation to the pattern of linkage disequilibrium observed at its breakpoint among the ancient standard lineages. However, several theoretical studies suggest that a newly arising inversion is likely to confer a fitness advantage in the presence of preexisting epistatic interactions (Kimura 1956; Wasserman 1968; Charlesworth and Charlesworth 1973; Charlesworth 1974; Alvarez and Zapata 1997).

Demography and selection are not mutually exclusive hypotheses and the two forces may in fact interact to produce even greater deviations than expected under either class of models (*cf.* Kaplan *et al.* 1991; Nordborg 1997; Slatkin and Wiehe 1998). Even if a long-standing epistatic interaction exists at the *In(2L)t* proximal breakpoint site in standard chromosomes, the recent increase

in *In(2L)t*'s frequency (Andolfatto *et al.* 1999) and demographic perturbations, such as population expansion, may have affected geographic patterns of variation for standard alleles at this locus. If both selection and demographic shifts have influenced patterns of variability, then the demographic history of *D. melanogaster* and *D. simulans* will have to be better understood before selection at any particular locus can be inferred.

We thank J. Comeron, R. Hudson, R. Kulathinal, M. Przeworski, S. C. Tsaur, and J. Wall for helpful discussions. R. Hudson and J. Wall provided computer programs. This manuscript was improved with comments from P. Awadalla, W. Eanes, an anonymous reviewer, and especially M. Przeworski. We thank Jean Gladstone for excellent technical assistance and Chung-I Wu for Australian and African fly lines. This research was supported by National Science Foundation grant DEB-9408869 and National Institutes of Health grant R01GM39355 to M.K. P.A. holds a Postgraduate Scholarship from the National Science and Engineering Council of Canada.

LITERATURE CITED

Aguadé, M., 1998 Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150:** 1079–1089.

Aguadé, M., 1999 Positive selection drives the evolution of the *Acp29AB* accessory gland protein in Drosophila. Genetics **152:** 543–551.

Alvarez, G., and C. Zapata, 1997 Conditions for protected inversion polymorphism under supergene selection. Genetics **146:** 717–722.

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

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

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, 1994 Evolutionary inferences from DNA variation at the *6-Phosphogluconate Dehydrogenase* locus in natural populations of *Drosophila*—selection and geographic differentiation. Genetics **136:** 155–171.

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

Bénassi, V., S. Aulard, S. Mazeau and M. Veuille, 1993 Molecular variation of *Adh* and *P6* genes in an African population of *Drosophila melanogaster* and its relation to chromosomal inversions. Genetics **134:** 789–799.

Bénassi, V., F. Depaulis, G. K. Meghlaoui and M. Veuille, 1999 Partial sweeping of variation at the *Fbp2* locus in a West African population of *Drosophila melanogaster*. Mol. Biol. Evol. **16:** 347–353.

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.

Charlesworth, B., 1974 Inversion polymorphism in a two-locus genetic system. Genet. Res. **23:** 259–280.

Charlesworth, D., and B. Charlesworth, 1973 Selection of new inversions in multi-locus genetic systems. Genet. Res. **21:** 167–183.

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., L. Brazier and M. Veiulle, 1999 Selective sweep at the *Drosophila melanogaster Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. Genetics **152:** 1017–1024.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triose-phosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **15:** 756–769.

Fu, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

Hale, L. R., and R. S. Singh, 1991 Contrasting patterns of genetic structure and evolutionary history as revealed by mitochondrial DNA and nuclear gene-enzyme variation. J. Genet. **70:** 79–89.

Hamblin, M. T., and C. F. Aquadro, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. Mol. Biol. Evol. **13:** 1133–1140.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of nucleotide sequence variation at the *glucose dehydrogenase* (*Gld*) locus in different populations of *Drosophila melanogaster*. Genetics **145:** 1053–1062.

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

Harada, K., S. I. Kusakabe, T. Yamazaki and T. Mukai, 1993 Spontaneous mutation rates in null and band-morph mutations of enzyme loci in *Drosophila melanogaster*. Jpn. J. Genet. **68:** 605–616.

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. J. Futuyma and J. Antonovics. Oxford University Press, Oxford.

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

Hudson, R. R., and N. F. 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. F. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

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., D. D. Boos and N. F. Kaplan, 1992a A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

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

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the *Superoxide Dismutase* (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Hudson, R. R., A. G. Sáez and F. J. Ayala, 1997 DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. Proc. Natl. Acad. Sci. USA **94:** 7725–7729.

Irvin, S. D., K. A. Wetterstrand, C. M. Hutter and C. F. Aquadro, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*: evidence for founder effects in New World populations. Genetics **150:** 777–790.

Kaplan, N., R. R. Hudson and M. Iizuka, 1991 The coalescent process in models with selection, recombination and geographic subdivision. Genet. Res. **57:** 83–91.

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

Kimura, M., 1956 A model of a genetic system which leads to closer linkage by natural selection. Evolution **10:** 278–287.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

Knibb, W. R., 1982 Chromosomal inversion polymorphism in *Drosophila melanogaster* II. Geographic clines and climatic associations in Australasia, North America and Asia. Genetica **58:** 213–221.

Kreitman, M., 1983 Nucleotide polymorphism at the *Alcohol Dehydrogenase* locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Labate, J. A., C. H. Biermann and W. F. Eanes, 1999 Nucleotide

variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans.* Mol. Biol. Evol. **16:** 724–731.

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

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

Maynard-Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable 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.

Nei, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

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

Slatkin, M., and T. Wiehe, 1998 Genetic hitchhiking in a subdivided population. Genet. Res. **71:** 155–160.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

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.

Tsaur, S. C., C. T. Ting and C. I. Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa,* of Drosophila: II. Divergence versus polymorphism. Mol. Biol. Evol. **15:** 1040–1046.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

Wasserman, M., 1968 Recombination-induced chromosomal heterosis. Genetics **58:** 125–139.

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

Communicating editor: A. G. Clark