Unusual Haplotype Structure at the Proximal Breakpoint of *In(2L)t* in a Natural Population of *Drosophila melanogaster*

Peter Andolfatto,* Jeffrey D. Wall† and Martin Kreitman*,†

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

Manuscript received September 23, 1998
Accepted for publication July 9, 1999

ABSTRACT

The existence of temporally stable frequency clines for *In(2L)t* in natural populations of *Drosophila melanogaster* suggests a role for selection in the maintenance of this polymorphism. We have collected nucleotide polymorphism data from the proximal breakpoint junction regions of *In(2L)t* to infer its evolutionary history. The finding of a novel *LINE*-like element near the *In(2L)t* breakpoint junction in sampled inverted chromosomes supports a transposable element-mediated origin for this inversion. An analysis of nucleotide variation in a Costa Rican population sample of standard and inverted chromosomes indicates a unique and relatively recent origin for *In(2L)t*. Additional *In(2L)t* alleles from three geographically diverse populations reveal no detectable geographic differentiation. Low levels of *In(2L)t* nucleotide polymorphism suggest a recent increase in the inversion's frequency in tropical populations. An unusual feature of our sample of standard alleles is a marked heterogeneity in levels of linkage disequilibrium among polymorphic sites across the breakpoint region. We introduce a test of neutral equilibrium haplotype structure that corrects both for multiple tests and for an arbitrarily chosen window size. It reveals that an ∼1.4-kb region immediately spanning the breakpoint has fewer haplotypes than expected under the neutral model, given the expected level of recombination in this genomic region. Certain features of our data suggest that the unusual pattern in standard chromosomes is the product of selection rather than demography.

INVERSION polymorphisms in the genus Drosophila are widely believed to be among the best examples of balanced polymorphisms. They have been used extensively as model systems for the study of the adaptive processes involved in the maintenance of genetic variation (reviewed in Krimbas and Powell 1992). *In(2L)t* is a common polymorphic inversion with stable frequency clines in natural populations of *Drosophila melanogaster*. While relatively rare in temperate climates (*i.e.*, <10% at latitudes above 35°), *In(2L)t* reaches frequencies of 40–60% in tropical populations of Australasia and Africa (Knibb 1982; Bénassi *et al.* 1993). The existence of parallel latitudinal clines across different continents and hemispheres offers compelling evidence that natural selection is maintaining inversion polymorphisms (Knibb 1982).

Population genetic models suggest that a possible advantage of inversions lies in their ability to suppress recombination in karyotypic heterozygotes and thus maintain favorable epistatic interactions between alleles at linked loci (Wasserman 1968; Dobzhansky 1970). Despite the evidence for selection on inversions based on biogeographical patterns (Knibb 1982) and cage experiments (Van Delden and Kamping 1991), surveys of molecular variation on standard and inverted chromosomes have found patterns that are generally consistent with historical and/or equilibrium neutral explanations (Ishii and Charlesworth 1977; Mukai and Voelker 1977; Nei and Li 1980; Strobeck 1983; Aquadro *et al.* 1986, 1991; Aguadé 1988; Bénassi *et al.* 1993; Rozas and Aguadé 1994).

In Drosophila, the linkage disequilibrium between selected sites and linked neutral variants predicted by epistatic or balancing selection models decays rapidly in the presence of gene conversion (Ishii and Charlesworth 1977; Strobeck 1983; Hudson and Kaplan 1988; Andolfatto and Nordborg 1998). So the effect of selection on linked neutral variation will not be detected if the rate of exchange between standard and inverted arrangements is too high. The rate of double crossover events and gene conversion has been estimated to be on the order of 10^{-4} to 10^{-5} per site per generation in the middle of inversions (reviewed in Ashburner 1989). In addition, gene conversion events between chromosomal arrangements are readily observed in nucleotide polymorphism data (Rozas and Aguadé 1994). Given such high rates of exchange, the pattern at linked neutral sites may not be particularly informative in such regions.

An alternative is to investigate the distribution of molecular variation within and between chromosomal arrangements at loci closely linked to the inversion breakpoints (Strobeck 1983; Andolfatto and Nord-

Corresponding author: Peter Andolfatto, Department of Ecology and Evolution, 1101 E. 57th St., University of Chicago, Chicago, Illinois 60637. E-mail: pandolfa@midway.uchicago.edu

borg 1998). Exchange between karyotypes is likely to be reduced at inversion breakpoints due to topological constraints on homologous pairing (Novitski and Braver 1954; Grell 1962; S. Hawley, personal communication). Two studies of loci closely linked to inversion breakpoints in Drosophila (Rozas and Aguadé 1994; Wesley and Eanes 1994) lead to different conclusions regarding the level of exchange between arrangements.

We have investigated nucleotide variation in a 5-kb region surrounding the proximal breakpoint of *In(2L)t*. Our *a priori* prediction is that the long-term maintenance of the inversion by selection will have led to the accumulation of a large number of fixed differences between karyotypic classes. Wesley and Eanes (1994) found no evidence for deviations from neutral equilibrium expectations at the breakpoints of *In(3L)P*, another common polymorphic inversion in *D. melanogaster* with frequency clines in nature similar to those of *In(2L)t*. One explanation is that selection does not often act long enough for the predicted signature of balancing selection to develop (Hudson *et al.* 1997). If instead *In(2L)t* is a young balanced or locally adapted polymorphism that has recently increased in frequency, we might observe a paucity of nucleotide variation among inverted alleles. By investigating sequences directly spanning the inversion breakpoint, our chances of observing the signature of selection, whether old or recent, will be increased due to the expected reduction in levels of genetic exchange between arrangements.

An unexpected feature of our sample of standard alleles is an unusually strong association among segregating sites close to the breakpoint. The presence of selection and recombination in a region can lead to considerable heterogeneity in levels of linkage disequilibrium (Hudson and Kaplan 1988; Hudson *et al.* 1997). We introduce a test based on Strobeck (1987) to detect subsets of the data that contain fewer haplotypes than expected under a neutral equilibrium model. Previous tests have not corrected for multiple tests nor for an arbitrarily chosen window size (Kirby and Stephan 1996; Bénassi *et al.* 1999); we show that this can have an important effect on the interpretation of the data.

MATERIALS AND METHODS

Localizing the 34A breakpoint: The *In(2L)t* proximal breakpoint (34A8-9) was first localized by *in situ* hybridization on polytene chromosomes (modified from Sniegowski and Charlesworth 1994) to two P1 clones (DS00576, DS01619). Subcloning of the shared regions of these two clones led to the recovery of an 8.4-kb *Sst*I fragment containing the 34A breakpoint (see Figure 1). This 8.4-kb region was randomly sheared by aspiration. Sheared fragments were repaired with Pwo polymerase (Boehringer Mannheim, Indianapolis). Size-selected fragments (0.5–1.6 kb) were ligated into a derivative of pZerO-2.1 (Invitrogen Inc., San Diego) and chemically transformed. Sequencing templates were prepared by PCR directly from colonies with standard M13-based primers, followed by polyethylene glycol (PEG) precipitation. Templates were sequenced with standard primers using a dRhodamine terminator cycle sequencing kit (Applied Biosystems Inc., Foster City, CA) and run on an ABI377XL automated sequencer. Sequences were analyzed with ABI sequence analysis v. 3.0 software. Contigs were managed with Sequencher v. 3.0 software. Inverted junction regions A/C and B/D (see Figure 1) were recovered by inverse PCR techniques (Ochman *et al.* 1993) and were sequenced as above. Sequences corresponding to the 8.4-kb proximal breakpoint and *In(2L)t* breakpoint junction regions have been deposited in GenBank under accession nos. AF172291–AF172316. An alignment has been deposited in the EMBL database under accession no. DS39454 (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/). The homologous 5-kb region for one line of *D. simulans* was sequenced for divergence estimates (GenBank accession no. AF172317).

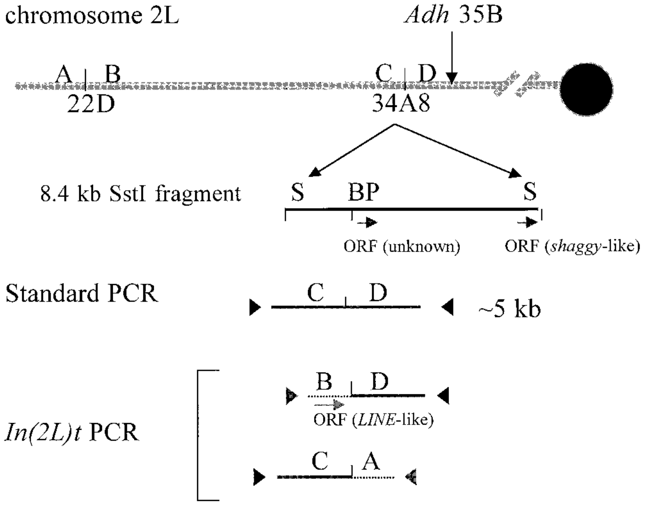


Figure 1.—Cloning and PCR strategy for the proximal breakpoint of *In(2L)t*. The proximal (34A8-9) breakpoint of *In(2L)t* is denoted by BP on an 8-kb *Sst*I (S) fragment. The position and orientation of putative open reading frames (ORFs) are denoted by arrows. A 5-kb region (C/D) from standard and inverted chromosomes was sampled for sequencing from inversion heterozygotes by selective PCR.

Population samples: Field collections were made from a single, large population of *D. melanogaster* in San Jose, Costa Rica. This tropical population was chosen because it was likely to have an appreciable frequency of *In(2L)t*. Wild-caught females were used to start isofemale lines. After several days of egg laying, genomic DNA was prepared from each original wild-caught female and PCR genotyped for *In(2L)t*. The use of specific primer pairs allowed the recovery of standard or *In(2L)t* alleles from inversion heterozygotes identified in the above screen. For the sampling of standard alleles, a 5-kb region spanning the proximal breakpoint (C/D in Figure 1) was PCR amplified from *In(2L)t* heterozygotes (primer C, GCCACGCCCCCATTCACTTAC; primer D, AATGCTTGGT GGCTTTGGAATGGT). These PCR-generated templates were sequenced directly with a set of forward and reverse primers. Similarly, for *In(2L)t* alleles, the two breakpoint junction regions (A/C and B/D in Figure 1) were amplified separately and sequenced from seven individuals (primer InA412, TTCGATCCACCGACAATCTGAAC; primer InB124, GTAC TTTCACTGTTTGCTGACGACGC). Inversion frequencies were also determined by PCR for three other populations (Florida City, FL; Yeppoon, Australia; and Zimbabwe, Africa) for comparison to previously published estimates. Australian and African isofemale lines were kindly provided by C.-I Wu. Primers used for karyotyping individual flies for *In(2L)t* are as follows: StC545, GACTCATTCTGCTTCGATCACTAAG; StD18, CTGTTCCCACCGCACAGAGTTGCCTGTC; InA151, TATTTTGGTGGCCTGTTTCAG. Expected PCR products are ∼500 bp for primer pair StC545/StD18 and ∼250 bp for primer pair StC545/InA151.

Polymorphism analyses: Tests of the neutral equilibrium model that compare polymorphism and divergence (*e.g.*, Hudson *et al.* 1987) could not be used here because of uncertainties in the alignment between *D. melanogaster* and *D. simulans*. The neutral mutation parameter $\theta = 4Nu$, where N is the effective population size of the species and u is the neutral mutation rate, was estimated from both π, the average pairwise difference per base pair (Tajima 1983), and S, the number of polymorphic sites in the sample (Watterson 1975). We use Tajima's D statistic (Tajima 1989) to characterize the skew in the frequency distribution of segregating mutations in our sample. Significance levels for Tajima's D and statistical tests described below were determined using coalescent simulations that condition on intermediate levels of recombination and the observed number of polymorphic sites in the sample (Hudson 1993). Similar simulations were used to test for the uniformity of polymorphic sites. Here, the statistic used was the length of the longest distance (in base pairs) between two consecutive segregating mutations. All simulations assumed a single panmictic Wright-Fisher population, unless otherwise specified. All simulations were run using modifications of a program kindly provided by R. Hudson.

Tree construction, *In(2L)t* historical frequency, and age: Bootstrapped parsimony phylograms were constructed using PAUP3.1.1 (Swofford 1993). To estimate an age for *In(2L)t*, we assume that the inversion has a unique origin (from a standard ancestor) and that the inverted region has been genetically isolated from standard chromosomes. Under a model in which balancing selection maintains the inversion, we assume the inversion immediately rose to a frequency f, where it has been maintained until the present. An alternative model assumes that the inversion is a strictly neutral polymorphism that has drifted to its present frequency.

If we assume the former model and that the inversion is at equilibrium, then θ and f can be estimated from the sample sizes of inverted and standard chromosomes (n_{i} and n_{s}, respectively) and the observed number of segregating sites within the inverted and standard chromosomes (S_{i} and S_{s}, respectively; Hudson and Kaplan 1986). Define $a_k = \Sigma_{j=1}^{k-1} j^{-1}$. Then, $E[S_{\mathrm{i}}] = \theta f a_{n_i}$ and $E[S_{\mathrm{s}}] = \theta(1 - f)a_{n_S}$. To estimate θ and f, the expectations are replaced by their observed values. This yields

$$\hat{f} = \frac{S_{\mathrm{i}}a_{n_{\mathrm{s}}}}{S_{\mathrm{s}}a_{n_{\mathrm{i}}} + S_{\mathrm{i}}a_{n_{\mathrm{s}}}} \tag{1}$$

and

$$\hat{\theta} = \frac{S_{\mathrm{s}}a_{n_{\mathrm{i}}} + S_{\mathrm{i}}a_{n_{\mathrm{s}}}}{a_{n_{\mathrm{i}}}a_{n_{\mathrm{s}}}}. \tag{2}$$

It is then straightforward to estimate the time to the most recent common ancestor of the inverted chromosomes by substituting (1) into

$$E[T_{\mathrm{MRCA}}] = 4\,N_{\mathrm{e}}f\,(1 - n_{\mathrm{i}}^{-1}). \tag{3}$$

This yields an estimate of the *minimum* age of the inversion. An *expected* age of the inversion is also calculated based on the average net number of differences between arrangements scaled to the expected divergence time between *D. melanogaster* and *D. simulans*. It should be noted that these estimates have very large variances.

We investigate by simulation whether levels of *In(2L)t* variation are compatible with assumed historical inversion frequencies under a neutral equilibrium model. We run 100,000 coalescent simulations with sample size n_{i} and population mutation parameter $\hat{\theta}f$. Here θ is estimated from (2), and we determine for which values of f the observation of S_{i} segregating sites is within the middle 95% of the simulated distribution of S_{i}.

Test of neutral equilibrium haplotype structure: We constructed a test to determine whether any subsets of consecutive segregating sites in our data contain fewer distinct haplotypes than would be expected under a neutral equilibrium model. Suppose we have a data set with n chromosomes and S segregating sites. 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 100,000 simulations using an infinite-sites, panmictic coalescent model conditional on n and S (Hudson 1993). We then calculate the proportion p_k of simulated data sets containing at least one stretch of S_k consecutive segregating sites with k or fewer haplotypes. This is equivalent to calculating the proportion of simulated data sets having S_k greater than or equal to the S_k observed in the data. Since choosing any value of k is arbitrary, we must correct for the implicit multiple tests involved. This corrected value P is determined from further coalescent simulations that compare the actual smallest p_k value with simulated smallest p_k values. Our test is similar to other tests that use the observed number of haplotypes as a test statistic (*e.g.*, Strobeck 1987; Fu 1996), but our test corrects for multiple windows and an arbitrarily chosen window size. Here window size is measured in segregating sites, not in base pairs, so our test shares affinities with commonly used "scan statistics" (*e.g.*, Karlin and Macken 1991).

A concern is the violation of the "infinitely many sites" assumption of our null model. In particular, mutations can sometimes overlap in actual data, especially in the presence of deletions. It is not clear whether "infinite-sites" simulations are conservative when the actual data has overlapping mutations. On one hand, more windows are considered in the simulations than in the actual data, which is conservative. On the other, missing information for those chromosomes with deletions might lead to fewer haplotypes than expected under the infinite-sites assumption. We believe that both effects are rather minor, but a higher degree of caution should be exercised when interpreting the relevant significance levels.

Two different versions of the *In(2L)t* data are considered. In-dels are likely to be governed by a different mutational process than single nucleotide polymorphisms (hereafter SNPs); however, they are included in simulations that condition on S since nothing is assumed about the underlying mutation rate, θ (Hudson 1993). Such information has been included in previous studies (*e.g.*, Hudson *et al.* 1994; Kirby and Stephan 1996; Cirera and Aguadé 1997). We have collapsed suspicious clusters of polymorphisms in our data into single events (complex mutations, denoted M in Figure 2). Version I includes SNPs and biallelic in-del polymorphisms, while Version II also includes overlapping mutational events.

All parameters used in simulations, including n, S, and C are listed in Table 1. In reality, the effective rate of recombination depends on the arrangement class. Our assumptions of panmixia and the same recombination rate for all individuals are unrealistic in this regard. Nevertheless, we consider it to be a reasonable simplifying assumption. In fact, an explicit demographic model of equilibrium-balancing selection based on estimates of the inversion's age and historic frequency yielded more significant P values (results not shown). Since individuals were not sampled randomly, both versions used a random sample based on the observed frequency (20.8%) of *In(2L)t*

TABLE 1

Summary of polymorphism data for the *In(2L)t* proximal breakpoint

	n^a	S	$C_{\mathrm{con}}{}^b$	θ^c	π	Tajima's D
Representative sample[d]	14	154	0.0003	0.0104	0.0125	1.00[e]
Standard alleles	11	146	0.0005	0.0100	0.0122	1.07
In(2L)t alleles	7	12	—	0.0011	0.0009	−0.79

C_{con} θ, π are given per site. In-del variation is included in calculations of θ and π.

[a] Number of alleles sampled.

[b] The population recombination rate for which our haplotype test (see Table 4) P is maximal.

[c] Watterson's (1975) estimate.

[d] Based on 21% *In(2L)t* frequency. Assuming other compatible *In(2L)t* frequencies did not significantly change these estimates.

[e] Tajima's test by simulation (see methods); $P > 0.10$ when $C = 0$; $P = 0.001$ when $C = C_{\mathrm{lab}}$.

chromosomes in the San Jose population. This sample included all 11 standard chromosomes as well as cr40i, cr44i, and cr46i (see Figure 2). Since *In(2L)t* varies in frequency among geographic locations, and the inferred historical frequency was estimated to be between 4 and 23% (see results), we also considered representative samples with zero to four inverted alleles.

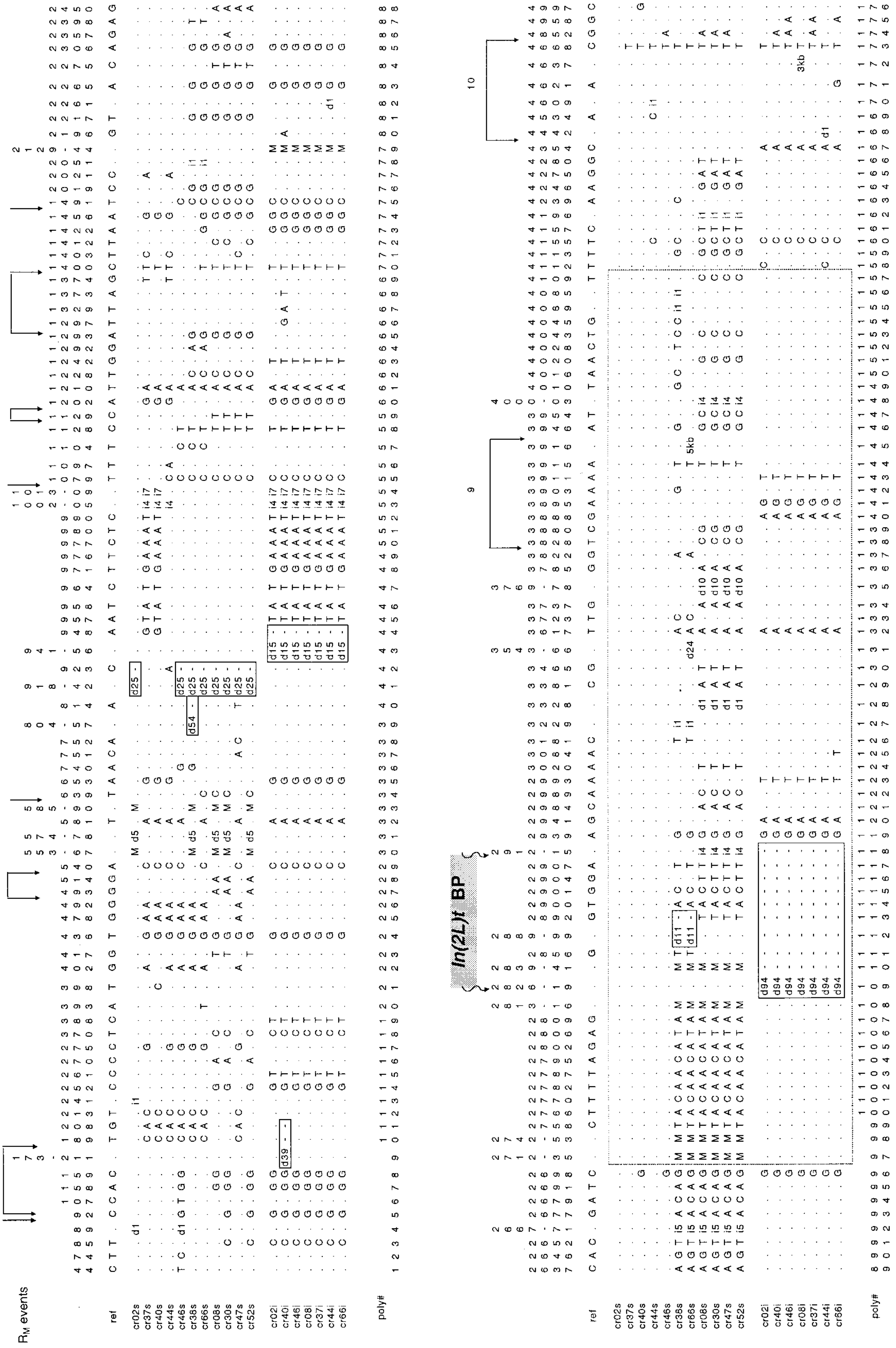
Recombination: The implications of recombination on the predictions of population genetic models are poorly understood. Perhaps as a result, investigators often conduct statistical tests with the assumption of no recombination. However, since there is evidence for recombination in our (and other) data, it is more realistic to assume $C > 0$ in simulations. In addition, ignoring the effect of recombination does not always lead to a conservative test (Wall 1999; results).

The use of estimators of the population recombination rate, $C = 4Nr$, on the basis of nucleotide polymorphism data (*e.g.*, Hudson 1987; Hey and Wakeley 1997) are problematic because they have large variances and are not necessarily conservative for our test. We consider two estimates of the recombination parameter to be informative. The first, C_{con}, is defined as the recombination rate for which the P value of our haplotype test is maximal. Assuming $C = C_{\mathrm{con}}$ makes our statistical test conservative with respect to our uncertainty over the true value of C. C_{lab} is a second estimate of $4Nr$ and assumes that $r = \rho$, where ρ is the per base pair, per generation rate of recombination obtained from laboratory measurements of the exchange of flanking markers (*cf.* Kliman and Hey 1993; Comeron *et al.* 1999). N is estimated as $\theta/4\mu$, where θ is estimated from the average nucleotide diversity at silent sites and noncoding DNA of loci in regions of intermediate recombination (Moriyama and Powell 1996). The neutral mutation rate, μ, is assumed to be 3×10^{-8} per site per generation, about twice that estimated for silent sites in Drosophila (Sharp and Li 1989; Fleischer *et al.* 1998). C_{lab} is not conservative in statistical tests. However, C_{lab} has the advantage of being independent of our polymorphism data set (and thus independent of neutral equilibrium assumptions) and it represents our best *a priori* guess at the true recombination rate in the chromosomal region studied.

The presence of an inversion in our sample makes the parameter C difficult to interpret. We expect that reduced exchange in inversion heterozygotes (*i.e.*, genetic isolation) will affect the overall levels of exchange. When calculating C_{lab} for use in simulations, we account for lack of recombination in *D. melanogaster* males and assume an absence of recombination in inversion heterozygotes. Thus, $C_{\mathrm{lab}} = (1 - 2q(1 - q))\ 2N\rho$, where q is the assumed inversion frequency. Note that the maximum effect of an inversion on the recombination rate among standard chromosomes is a factor of two (assuming an inversion at 50% frequency and no recombination in inversion heterozygotes).

Analysis of other data sets: To demonstrate the effect that an arbitrarily chosen window size can have on the likelihood of a data set, we assess the significance of nonneutral haplotype structure reported in additional studies of *D. melanogaster* populations: the *white* locus data of Kirby and Stephan (1996), the *Acp70A* locus data of Cirera and Aguadé (1997), and the *vermilion* locus data of Begun and Aquadro (1995). All biallelic mutations were considered for the analysis of *white* (Figure 1 of Kirby and Stephan 1996). We excluded site 383, a multiple hit, in the *Acp70A* data (see discussion). For the *vermilion* data, each population sample was considered separately. Sites 1031 and 2532 (which are not biallelic) were

Figure 2.—Summary of polymorphic sites found in 5012 bp at the proximal breakpoint of *In(2L)t*. The 11 standard and 7 inverted chromosomes are labeled s and i, respectively. A dot represents identity with the reference sequence. The nucleotide position in base pairs is indicated above the reference sequence; the polymorphic site number is indicated below. Insertion/deletion polymorphisms are denoted by i and d, respectively. Complex polymorphisms and denoted by M. A small polymorphic homopolymer run of thymidines (length 9–16) was found between positions 2287 and 2302 but was not included because its length could not always be reliably determined. Two transposable element polymorphisms were also found: cr66s contained an ∼5-kb Burdock element at position 3916 and cr08i contained an ∼3-kb insertion at position 4637, which appears to be a composite of a P element and an unidentified segment of DNA. Inferred recombination events (R_{M}) are indicated with straight arrows. Two curved arrows indicate the position of the *In(2L)t* breakpoint. The dotted box (encompassing polymorphisms 97–157) corresponds to S_5 (see Table 3), a 1.4-kb region containing 60 polymorphisms in a representative random sample (*In(2L)t* = 21%) that has fewer haplotypes than expected under a neutral equilibrium model, assuming $C = C_{\mathrm{lab}}$.


R_M events
1
2
3
4
5
6
7
8
9
10
In(2L)t BP
ref
cr02s
cr37s
cr40s
cr44s
cr46s
cr38s
cr66s
cr08s
cr30s
cr47s
cr52s
cr02i
cr40i
cr46i
cr08i
cr37i
cr44i
cr66i
poly#
5kb
3kb

excluded. P values of haplotype tests are reported for $C = C_{con}$ and $C = C_{lab}$. C_{lab} is estimated for each locus, as above, using estimates of ρ for each locus (Comeron *et al.* 1999).

RESULTS

Polymorphism summary: Figure 2 summarizes 176 polymorphic sites found in the sampled 5-kb of region C/D (see Figure 1). The data set contains 131 biallelic single nucleotide polymorphisms (SNPs). In addition, there are 27 biallelic insertion-deletion polymorphisms, denoted by i, d, or M. Complex polymorphisms are denoted by M; they are either multiple insertions (or deletions) that are very close to each other, or one or more nucleotide mutations associated with an insertion or deletion. The spatial distribution of insertion-deletion variation does not appear to be different from that for SNPs and does not significantly affect Tajima's D statistic (results not shown). The *In(2L)t* breakpoint occurs between positions 2826 and 2919, where all sampled inverted chromosomes ($n = 18$) contain a 94-bp deletion.

Inversion frequencies and uniqueness of origin: Estimates of *In(2L)t* frequencies assessed by PCR assay in Yeppoon and Florida City samples were 22.9% ($n = 43$) nd 25.0% ($n = 100$), respectively. These estimates are similar to those based on the cytology of geographically proximate populations (Mettler *et al.* 1977; Knibb 1982). *In(2L)t* frequencies for San Jose and Zimbabwe (Harare and Sengwa) were estimated to be 20.8 ($n = 100$) and 58.2% ($n = 98$), respectively. To confirm that all *In(2L)t* chromosomes share the same breakpoint, a number of extracted *In(2L)t* chromosomes (Costa Rica, $n = 10$; North Carolina, $n = 7$; Zimbabwe, $n = 10$) were assayed by PCR. In all cases there was agreement between the cytology and PCR results consistent with a single mutational origin for *In(2L)t*. Sequence from 18 *In(2L)t* chromosomes (San Jose, $n = 7$; Florida City, $n = 3$; Yeppoon $n = 4$; and Zimbabwe, $n = 4$) revealed that all have identical breakpoints.

Recombination within and between chromosomal arrangements: For the chromosomal band 34A8-9, the laboratory estimate of the crossover rate is 2.9×10^{-8} per site per generation in females (Comeron *et al.* 1999). Assuming an *In(2L)t* frequency of 20.8%, no recombination in inversion heterozygotes and a species population size of 10^6 (see methods), C_{lab} is 0.019 per site per generation. There is evidence for a considerable amount of recombination in our sample of standard chromosomes. The positions of 10 inferred minimum recombination events (Hudson and Kaplan 1985) in Figure 2 are denoted by arrows. Adding inverted chromosomes to the analysis revealed no additional recombination events.

Despite the large number of informative polymorphisms, there is almost no evidence for genetic exchange between arrangements in our data. A possible exception is site 174 (Figure 2), which may be a multiple hit. This observation is consistent with the finding of suppressed exchange between standard and inverted classes at the breakpoint of the inversion (Novitski and Braver 1954; Grell 1962). We expect that genetic exchange between chromosomal arrangements will tend to decrease the number of fixed differences between arrangements and increase the number of shared polymorphisms. Of the 120 informative polymorphic sites, we observed 12 fixed differences between arrangements and just 1 shared difference. This pattern is in marked contrast to patterns observed at two additional loci linked to *In(2L)t* (summarized in Table 2). These two loci, *Adh* and *Fbp2*, are further from the breakpoints where the rate of exchange between arrangements is likely to be higher.

History of *In(2L)t*: A bootstrapped parsimony phylogram for 2.3 kb of sequence (positions 1491 to 3822) immediately spanning the inversion breakpoint is shown in Figure 3. All mutations in this segment are consistent with a single genealogical tree (*i.e.*, there are no inferred recombination events in the segment). From the tree, we can infer that the inversion is relatively recently derived from one of two distinct haplotype classes of standard chromosomes. Trees based on three other recombination-free segments (positions 189–492, 639–1020, and 3946–4354) all produce similar topologies: *In(2L)t* alleles form a distinct cluster that is closely related to one or more standard haplotypes.

In total, 12 biallelic segregating polymorphisms were found in the 7 sampled *In(2L)t* alleles. Of these, 9 are SNPs (see Figure 2). Estimates of θ and π for the *In(2L)t* alleles are 11 and 7%, respectively, of those for standard alleles (Table 1). Assuming the population is at equilibrium, this level of diversity is consistent with a low (*i.e.*, ~10%) average historical frequency.

In(2L)t appears to be recently derived relative to standard lineages and the *D. melanogaster*-*D. simulans* divergence time. We estimate the $E[T_{MRCA}]$ of our sample of *In(2L)t* alleles to be $\sim 0.3\ N_e$ generations. This roughly corresponds to ~100,000 years, assuming an $N_e \sim 3 \times 10^6$ (Kreitman 1983) and 10 generations per year. Our estimate does not change significantly over a wide range of assumed inversion frequencies. However, since *In(2L)t* is rare relative to the standard alleles, it is more often heterozygous and thus more susceptible to forces such as background selection (Charlesworth *et al.* 1993) and selective sweeps (Maynard Smith and Haigh 1974). These forces may lead to an underestimate of the historical frequency of *In(2L)t* (and thus the $E[T_{MRCA}]$). However, the magnitude of their effect cannot be meaningfully estimated without *a priori* knowledge of the inversion's true historical frequency. As an illustration, the resulting reduction in recombination relative to the standard class ($\sim q^2/(1 - q)^2$) will be greater than 350-fold for an inversion at 5% frequency, and only 9-fold for an inversion at 25%.

TABLE 2

Data on genetic exchange between arrangements at loci linked to inversions

Region	Cytological position	Standard π	Inverted π	Informative poly[a]	Fixed: shared[b]
In(2L)t breakpoint[c]	34A8-9	0.0122	0.0009	120	12:1
Adh-S and *Adh-dup*					
(Zimbabwe+World)[d]	35B3	0.0069	0.0029	70	0:20
(Ivory Coast)[e]		0.006	0.005	13	0:11
(Spain)[f]		0.0028	0.0046	14	0:6
Fbp2 (Ivory Coast)[e]	30B1-12	0.0072	0.0068	15	0:11
In(3L)P breakpoints[g]	63B8-11 and 72E1-2	0.0060	0.0003	42	8:0
Hsp83[h]	63C1	0.0065	0.0009	8	2:0
Est-6[h]	69A1-5	0.0162	0.0200	30	0:17
Est-5 D. pseudoobscura[i]	23 (near SR BP 22)	0.0123	0.0013	22	7:1
rp49 D. subobscura[j]	Near BP of O_{ST}/O_{3+4}	0.0062	0.0079	53	2:12

[a] The number of polymorphisms in the sample excluding singletons.

[b] The ratio of fixed polymorphisms between arrangements to the number shared between arrangements.

[c] Present study, including insertion-deletion variation, Table 1.

[d] S.-C. Tsuar, unpublished data; the world sample is that of Kreitman (1983).

[e] Data from Bénassi *et al.* (1993); restriction site data.

[f] Data from Aguadé (1988); restriction site data.

[g] Data from Wesley and Eanes (1994).

[h] Data from Hasson and Eanes (1996).

[i] Data from Rozas and Aguadé (1994).

[j] Data from Babcock and Anderson (1996).

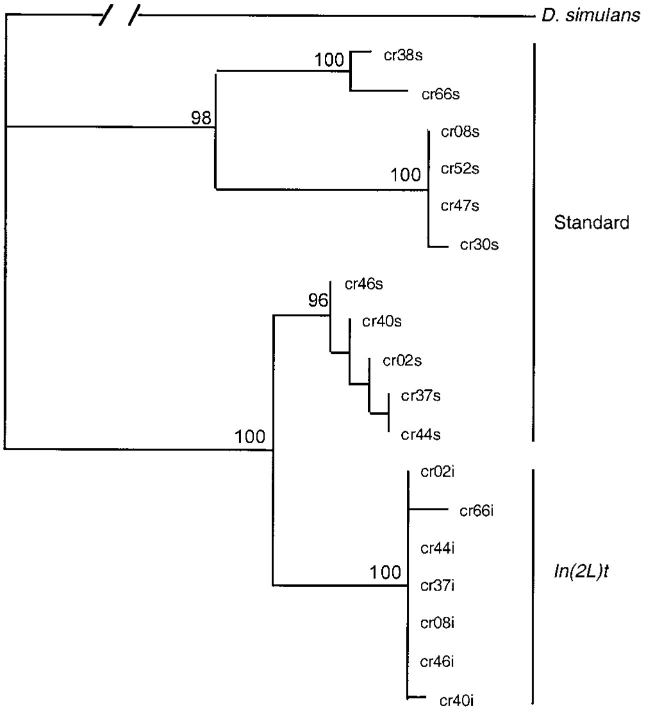


Figure 3.—Bootstrapped parsimony phylogram for a 2.3-kb region immediately spanning the *In(2L)t* breakpoint (positions 1491 to 3822 in Figure 2). The tree indicates that *In(2L)t* is relatively recently derived from one of two distinct standard haplotype classes. Numbers on nodes indicate the percentage of 1000 bootstrap replicates that supported the node.

A second estimate of the age can be obtained from the number of fixed differences between *In(2L)t* and standard lineages: the number (9) is small compared to the average pairwise divergence between standard chromosomes (48.5). This again suggests that *In(2L)t* is recently derived relative to standard lineages. We estimate the *expected* age of the inversion (~160,000 years) to be ~4% of the net divergence time for *D. melanogaster* and *D. simulans.*

***In(2L)t* nucleotide diversity and frequencies in natural populations:** Several features of the data suggest that *In(2L)t* is not at equilibrium. First, Tajima's *D* for the inverted class is negative (though the frequency spectrum is not significantly skewed). This finding is interesting given that Tajima's *D* is significantly positive for the random sample. Second, *In(2L)t* chromosomes show no segregating variation in a region of elevated θ (positions 2637 to 3020, Figure 2). These observations suggest a recent origin and rapid increase in the inversion's frequency.

Evidence that *In(2L)t* has recently increased in frequency can be inferred from the geographic distribution of *In(2L)t* variation and the relative levels of diversity for standard and inverted chromosomes. We sequenced an ~0.8-kb region spanning the C/A breakpoint junction of 11 additional *In(2L)t* chromosomes (positions 2001 to 2819, Figure 2). Table 3 summarizes polymorphic variation found in this subregion for 18 *In(2L)t* alleles from four geographically diverse populations. Permutation tests (Hudson *et al.* 1992) failed to

TABLE 3

In(2L)t variation in four population samples

	Position[a]			
Line	2195	2261	2449	2764
ref	G	—	G	G
cr02i	—	—	—	—
cr08i	—	—	—	—
cr37i	—	—	—	—
cr40i	A	—	—	—
cr44i	—	d1	—	—
cr46i	—	—	—	—
cr66i	—	—	—	—
fc84i	—	—	—	—
fc95i	—	—	—	—
fc100i	—	—	—	—
y01i	—	—	—	—
y07i	A	—	—	—
y11i	—	—	—	—
y25i	—	—	—	—
zh20i	—	—	A	A
zh29i	—	—	—	—
zs30i	—	—	—	A
zs53i	—	—	—	—

[a] The region sequenced corresponds to positions 2001 to 3068 in Figure 2. Labels of alleles correspond to the following populations: cr, San Jose, Costa Rica; fc, Florida City, FL; y, Yeppoon, Australia; zh and zs, Harare and Sengwa Wildlife Reserve, Zimbabwe.

detect geographic differentiation of *In(2L)t* alleles ($P > 0.4$). Using coalescent simulations, we estimate that the level of *In(2L)t* nucleotide diversity is inconsistent with equilibrium inversion frequencies greater than ~23% (assuming $C = C_{\text{lab}}$). Thus, the inversion is not significantly lacking in variation given its sampled frequencies in Costa Rica, Yeppoon, or Florida City. Under the most conservative and unrealistic assumption of no recombination, *In(2L)t* diversity is inconsistent with frequencies above ~35%. The Zimbabwe sample well exceeds this upper limit with an *In(2L)t* frequency of 58.2%. This sample is actually composed of two populations that had similar *In(2L)t* frequency estimates: Sengwa Wildlife Reserve (61.4%, $n = 44$) and the capital, Harare (55.5%, $n = 54$).

Heterogeneity in nucleotide polymorphism: Sliding-window profiles of nucleotide diversity (Figure 4) reveal considerable heterogeneity in levels of polymorphism across region C/D. For example, the large mutation-free stretch centered at ~1600 in Figure 4B (found between positions 1491 and 2019 in Figure 2) is highly unusual ($P < 0.004$ for all C) assuming a uniform mutation rate. Nonuniformity of mutations could arise from heterogeneity in selective constraint across the region. A comparatively large peak of intraspecific polymorphism is observed at the inversion breakpoint (approximately position 2600 in Figure 4A). Interestingly, most of this variation is distributed *between* two major haplotypic classes and not within them (Figure 4B), reflecting the strong linkage disequilibrium observed between sites close to the breakpoint (see Figure 2). The average pairwise divergence between haplotypic classes is ~10% at its highest point. This pattern is reminiscent of that observed between Fast and Slow allozyme classes at the *Adh* locus (Kreitman and Hudson 1991), a polymorphism believed to be maintained by balancing selection. The divergence between *Adh*-Fast and *Adh*-Slow allelic classes is between 6 and 7% at its highest point using a comparable window size. The average divergence between *D. melanogaster* and *D. simulans* at silent sites is ~9% (Takano 1998).

There seem to be fewer haplotypes at the breakpoint than expected under a neutral equilibrium model (given the sample size, the number of mutations, and the expected recombination rate). Unfortunately, Figures 4A and 4B are tricky to interpret: for one, the division of standard chromosomes into two "haplotypic classes," as done in Figure 4B, is *post hoc*. In addition, the chosen window size is arbitrary. Given evidence for heterogeneity in constraint in the region, Figure 4 could be misleading since each window may not contain an equivalent number of "neutral" polymorphisms. In general, it is unclear how to test whether an arbitrarily defined window is unusual *after* having examined the data. To address this difficulty, the test described below corrects for a *post hoc* choice of window size.

Test of neutral equilibrium haplotype structure: While there is evidence for recombination in our data (*i.e.*, 10 minimum inferred recombination events), there are also strong associations among sites over a considerable distance spanning the *In(2L)t* breakpoint (see positions 1491–4109 in Figure 2). Although this observation seems unlikely under the null model, closely linked sites do have correlated histories in neutral genealogies with intermediate recombination (Griffiths 1981). For this reason, tests based on the permutation of sites (*e.g.*, Leicht *et al.* 1993; Kirby and Stephan 1996), which assume sites are exchangeable, are not appropriate in this context. Instead, we tested the unusual features of our data by coalescent simulations with recombination. Table 4 shows p_k values from simulations run on the two versions of the $n = 14$ data (see methods). For example, Figure 2 shows 12 consecutive segregating sites (sites 97–108) that define only two haplotypes. To test how unusual this pattern is, 100,000 replicates were run using a panmictic, no recombination coalescent model that conditioned on $n = 14$ and $S = 154$. The proportion (p_k) of simulated runs containing 12 or more consecutive segregating sites that define only two haplotypes was $p_2 = 0.048$(see Table 4, version I, $C = 0$). The p_k values for each S_k, corrected for multiple windows, are reported for $C = C_{\text{con}}$ and $C = C_{\text{lab}}$. The

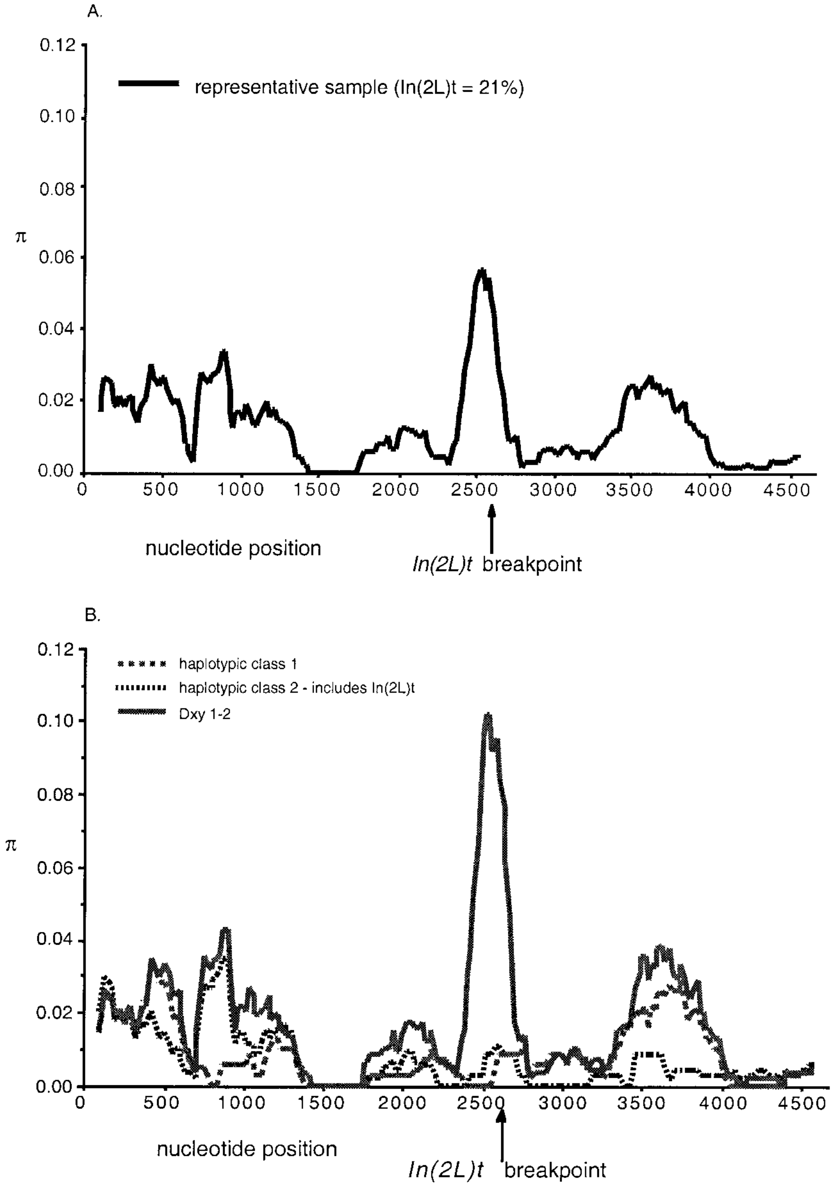


Figure 4.—(A) Sliding-window profile of nucleotide diversity (π) for a representative random sample of alleles (*In(2L)t* frequency = 21%). A peak of polymorphism is revealed at the inversion breakpoint, indicated by an arrow. The window size is 200 bp; the increment is 10%. Nucleotide positions do not correspond exactly to those in Figure 2 since gaps have been excluded. (B) The distribution of nucleotide diversity within and between haplotype classes. This analysis reveals that most of the diversity at the breakpoint is distributed *between* (solid line) and not *within* (dotted lines) the two haplotypic classes. Class 1 (defined arbitrarily) includes lines cr08s, cr30s, cr47s, cr52s, and cr66s; class 2 includes the remaining standard and *In(2L)t* alleles.

most unusual S_k for our data is for $S_5 = 50$ (or $S_5 = 60$ for Version II, Table 4).

Final P values that correct for a *post hoc* choice of window size are shown in Table 4. Each of these values reflect median P values for all possible random samples with three *In(2L)t* alleles (reflecting the sampled *In(2L)t* frequency of 20.8%). Since *In(2L)t* nucleotide diversity levels are consistent with frequencies of between ~4 and 23%, we conducted additional tests on random samples corresponding to *In(2L)t* frequencies of between 0 and 27%. Most of the tests on Version I of the data were not significant when $C = C_{\mathrm{con}}$. Version II of the data, which includes polymorphisms that overlap with the 94-bp deletion of *In(2L)t* alleles, is significant when $C = C_{\mathrm{con}}$ for samples that included between zero and three *In(2L)t* alleles.

Figure 5 shows P *vs.* C for Version I of the data. Note that $C = 0$ is not conservative and that P decreases (almost) monotonically for $C > C_{\mathrm{con}}$. The pattern in Figure 5 (an initial increase in P for low values of C followed by a steady decrease as C increases) is a general feature of our test (results not shown). When running simulations with our *a priori* estimate of the recombination rate (C_{lab}, corrected for the inversion's effect), all tests were highly significant ($P < 0.001$) for both versions of the data, including samples with zero to four *In(2L)t* alleles.

Multiple tests, recombination, and haplotype structure at other loci: Correcting for multiple tests and an arbitrary window size can have an impact on the significance of the data. For example, the most unusual subset of the *Acp70A* data (Cirera and Aguadé 1997) is a window of 19 segregating sites with three haplotypes ($p_3 = 0.064$, $C = C_{\mathrm{con}}$). By our test (Table 5), the corrected $P = 0.130$ ($C = C_{\mathrm{con}}$). For the *vermilion* data of Begun and Aquadro (1995), the lowest p_k is lower than

TABLE 4

Summary of haplotype tests on the *In(2L)t* proximal breakpoint

Data set version		k (no. haps)	S_k^a	Sites^b	p_k^c $C = C_{\text{con}}$	p_k^c $C = C_{\text{lab}}^d$	P^d $C = C_{\text{con}}$	P^d $C = C_{\text{lab}}^d$
I.		2	12	97–108	0.0602	0.0163		
		3	20	89–108	0.0576	0.0071		
		4	23	97–130	0.1400	0.0235		
	p_{min}	5	50	97–157	0.0199	0.0003	0.0645	_0.0000_*
		6	60	87–157	0.0331	0.0009		
		7	64	83–157	0.0851	0.0055		
II.		2	12	97–108	0.0716	0.0226		
		3	20	89–108	0.0675	0.0101		
		4	33	97–130	0.0377	0.0022		
	p_{min}	5	60	97–157	0.0088	0.0001	_0.0323_	_0.0000_*
		6	70	87–157	0.0178	0.0004		
		7	74	83–157	0.0525	0.0028		

The p_k and P values reflect the median of all possible samples including three (*i.e.*, 21%) *In(2L)t* alleles. Details of the test, including versions of the data and recombination estimates, are discussed in the text. Boxed rows correspond to the window size for which p_k is minimal. Significant tests are indicated in italic type. Version I of the data includes only biallelic SNPs and insertion-deletion polymorphisms. Version II includes polymorphic sites 9–11, 39–44 and 109–118 (see Figure 2). $^*P < 10^{-5}$.

[a] The largest window of segregating sites with k haplotypes.

[b] Polymorphic site numbers correspond to those in Figure 2.

[c] One-tailed probabilities, corrected for multiple tests.

[d] One-tailed probabilities, corrected for multiple tests and for the *post hoc* choice of window size.

[e] An *a priori* estimate of the local recombination rate (see methods).

0.05 for all non-African populations when $C = C_{\text{con}}$ (Beijing, California, Ecuador, and Taiwan). Two of these populations (Ecuador and Taiwan) are no longer significant when correcting for the *post hoc* choice of window size under conservative recombination (P = 0.064 and P = 0.091, respectively, Table 5).

The assumed rate of recombination also has a large effect on our interpretation of the data (Table 5). For example, under conservative recombination the *white* locus (Kirby and Stephan 1996) shows no sign of unusual haplotype structure ($P = 0.903$, $C = C_{\text{con}}$). However, highly significant nonneutral haplotype structure is detected for the *white* data set when we assume $C = C_{\text{lab}}$ ($P = 0.001$). As expected, highly significant haplotype structure is also detected in the *Acp70A* and non-African *vermilion* samples when $C = C_{\text{lab}}$ ($P \leq 0.004$). Interestingly, significant haplotype structure is not detected in the two African samples of *vermilion*, even when $C = C_{\text{lab}}$ (P = 0.817 and 0.270 for Kenya and Zimbabwe, respectively). Thus, these two samples are in accordance with the predictions of a neutral equilibrium model under expected levels of recombination (*i.e.*, $C = C_{\text{lab}}$).

Open reading frames: A scan of 8.4 kb of the proximal breakpoint sequence (see Figure 1) with open reading frame (ORF) Finder (http://www.ncbi.nlm.nih.gov) revealed an ~200-amino-acid open reading frame between positions 7701 and 8343. A search of the protein databases revealed that this putative exon encodes a *shaggy*-like protein kinase. The position and direction of this exon is indicated in Figure 1. This putative exon was not included in our polymorphism study. GRAIL v. 1.3 software (http://compbio.ornl.gov) identified a second putative ORF (Figure 1) between positions 3078 and ~3227. This putative ORF is oriented in the same direction as the *shaggy*-like ORF and contains a helix-loop-helix dimerization motif. It is unknown whether

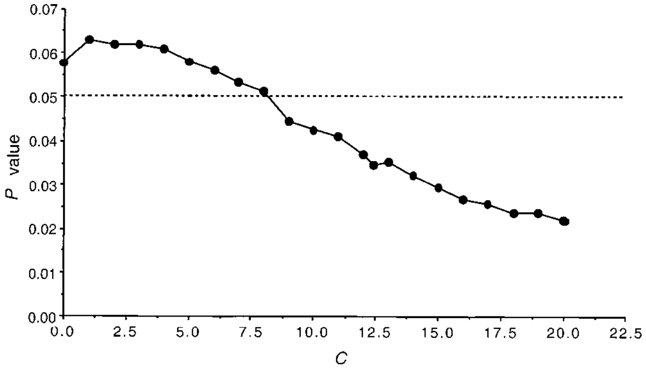


Figure 5.—Relationship between P value and $C = 4Nr$ for a representative random sample of alleles based on an *In(2L)t* frequency of 21%. Our *a priori* estimate of the population recombination rate corrected for inversion frequency. C_{lab} = 90. The general behavior of our statistic P with respect to C is robust to different representative samples of our data and the three additional data sets analyzed.

TABLE 5

Haplotype tests on *white*, *Acp70A*, and *vermilion*

	P	
Data set	$C = C_{con}$	$C = C_{lab}$
white	0.903	*0.001*
Acp70A	0.130	*0.004*
vermilion		
Beijing	*0.007*	*0.000**
California	*0.016*	*0.000**
Ecuador	0.064	*0.000**
Taiwan	0.091	*<0.001*
Kenya	0.996	0.817
Zimbabwe	0.927	0.270

Significant tests are indicated in italic type. *$P < 10^{-5}$.

or not these ORFs are contained in a single gene. The proximity of this putative coding region to the *In(2L)t* breakpoint raises the possibility that the inversion may affect patterns of gene expression. This finding would not be unprecedented in the literature (*cf*. Wesley and Eanes 1994).

A scan for ORFs in the *In(2L)t* B/D junction region revealed a 93-amino-acid open reading frame oriented toward the breakpoint that encodes the 3′-end of a novel putative reverse transcriptase of the *LINE* family of retroelements (Dinocera and Sakaki 1990). It is likely that this element was involved in *In(2L)t*'s formation since it lies within 0 and 20 bp of the B/D breakpoint junction. The element is present in all sampled inverted chromosomes ($n = 18$) and absent from all sampled standard chromosomes ($n = 11$). The element was also absent in a larger ($n = 24$), geographically diverse (including Miami, Yeppoon, and Zimbabwe) sample of standard chromosomes. Details of this novel transposable element will be described elsewhere.

DISCUSSION

Origin and age of *In(2L)t*: A *LINE*-like retrotransposon is found immediately at one of the breakpoint junctions. This element is absent in standard chromosomes but fixed in *In(2L)t* chromosomes, suggesting a transposable element-mediated origin for the inversion. This finding supports other recent work indicating a role for transposable elements in the formation of naturally occurring chromosomal rearrangements (Lyttle and Haymer 1992; Mathiopoulos *et al.* 1998).

Patterns of polymorphism and divergence between chromosomal arrangements in the San Jose population sample suggest that *In(2L)t* has a unique and relatively recent origin. Our estimate of *In(2L)t*'s age, ∼160,000 years old (or ∼1.6 million generations), is similar to the age reported for *In(3L)P* (Hasson and Eanes 1996). The inverted sample of alleles has relatively low levels of nucleotide diversity and divergence from standard alleles. A survey of three other geographically diverse populations revealed neither significantly more variation nor evidence for geographic differentiation of *In(2L)t* alleles. Thus, despite their wide geographic distributions and high frequencies in tropical populations (reviewed in Lemeunier and Aulard 1992), *In(2L)t* and *In(3L)P* are not long-lived balanced polymorphisms.

An alternative to the "long-lived balanced polymorphism" scenario is that *In(2L)t* has very recently increased in frequency in some populations due to selection. If so, the inverted class should have either reduced levels of variation, as observed in one allelic class at the *Sod* locus of *D. melanogaster* (Hudson *et al.* 1994), or a sharply negative skew in the frequency spectrum of segregating mutations (Braverman *et al.* 1995). The moderate number of fixed differences between karyotypes does suggest that *In(2L)t* is young relative to standard lineages. However, while Tajima's *D* is negative, there are probably too few segregating sites for us to have power to detect selection on *In(2L)t* based on the frequency spectrum of mutations (also true of *In(3L)P*; see Wesley and Eanes 1994).

***In(2L)t* is not at equilibrium in natural populations:** A recent increase in the frequency of *In(2L)t* can be inferred by considering relative levels of nucleotide diversity in standard and inverted arrangement classes (Tables 1 and 2) as well as world-wide inversion frequency estimates (Knibb 1982; Bénassi *et al.* 1993; this study). Our simulations indicate that levels of *In(2L)t* diversity are consistent with historical frequencies of between 4 and 23% when we assume $C = C_{lab}$ (or 3 and 35% when we assume $C = 0$). This implies that *In(2L)t* is not at equilibrium in populations with *In(2L)t* frequencies that exceed 23–35%. Examples of such populations include those in northern Australia, New Guinea, southern Japan (Knibb 1982), West Africa (Bénassi *et al.* 1993), and Zimbabwe (this study). Despite extensive biogeographical and experimental evidence for selection on inversions (reviewed in Krimbas and Powell 1992), our results represent the first evidence from nucleotide polymorphism data that inversions are not at neutral equilibrium in tropical populations. Possible explanations for a recent increase in frequency of *In(2L)t* include founder and/or hitchhiking events associated with the inversion. However, the lack of geographic differentiation of *In(2L)t* alleles and the inversion's high frequency in multiple populations suggests that its high frequency is not an isolated phenomenon. Since the expected time to equilibrium is relatively long (∼4*Nf* generations), *In(2L)t* could be a locally adaptive or balanced polymorphism that is too young to be at equilibrium.

Genetic exchange between arrangements is suppressed at inversion breakpoints: Our data suggest that exchange between arrangements, including gene conversion, is significantly suppressed at breakpoint regions relative to other regions. This is manifested by the lack of shared polymorphism between arrangements (Table 2) and the low diversity of *In(2L)t* and *In(3L)P* alleles (relative to standard alleles) at the junction region. We find no evidence for genetic exchange between karyotypic classes despite the large number of informative sites in the region sequenced (with the possible exception of site 174, Figure 2). In contrast, patterns of variation at *Adh* and *Fabp2* are consistent with considerable levels of between-karyotype exchange at these loci. Patterns of nucleotide polymorphism at multiple loci linked to *In(3L)P* lead to a similar conclusion (Table 3).

The distribution of polymorphism within and between arrangements at *rp49* is significantly different from that observed at both *In(2L)t* and *In(3L)P* breakpoints (Table 2, two-tailed $P < 0.01$ and $P < 0.04$, respectively, by a Fisher's exact test). *Est-5* of *D. pseudoobscura* (Table 2) shows a pattern more consistent with the patterns observed at *In(2L)t* and *In(3L)P*. The *rp49* data may be unusual due to the complex nature of the associated inversion system in *D. subobscura* (Rozas *et al.* 1999). With the exception of *rp49*, the data suggest that the rate of exchange between karyotypes at inversion breakpoints is likely to be lower than the reciprocal of the effective population size of the species (Ishii and Charlesworth 1977) and is not constant across the inverted region as is assumed in some recent models (*e.g.*, Navarro *et al.* 1997).

Unusual haplotype structure at the *In(2L)t* breakpoint: An unexpected feature of our data is a marked heterogeneity in levels of linkage disequilibrium across the sequence (Figure 2) among standard chromosomes. Two deeply diverged standard arrangement haplotypes exhibit strong associations among sites close to the inversion breakpoint. Yet there is ample evidence for recombination in other regions of the data set.

Our statistical test reveals that an ~1.4-kb region immediately spanning the *In(2L)t* breakpoint (Figure 2) has too few haplotypes to be compatible with a neutral equilibrium model under the expected level of recombination (*i.e.*, C_{lab}; see Table 4). It also suggests that the haplotype structure observed at *In(2L)t* is more extreme than at two other loci (*i.e.*, *white* and *Acp70A*, Table 5) where unusual haplotype structure had previously been reported (Kirby and Stephan 1996; Cirera and Aguadé 1997). An advantage of our test is that it can identify unusual regions that are not apparent when initially viewing the data. The level at which the null hypothesis could be rejected depends on both the amount of information included (*i.e.*, with different versions of the data) and the assumed recombination rate.

Unfortunately, the true recombination rate is unknown. Simulations run with C_{con} do not always show a significant deficiency in the number of haplotypes at the *In(2L)t* breakpoint. However, on the basis of comparisons of physical and genetic maps (Kliman and Hey 1993; Comeron *et al.* 1999) and estimates of the effective population size (Kreitman 1983) of *D. melanogaster*, it seems likely that C_{con} is a severe underestimate of the true population recombination rate. As an illustration, C_{lab}, our best *a priori* estimate of C in this chromosomal region (corrected for the presence of the inversion), is >50-fold higher than C_{con} and ~10-fold higher than the C corresponding to $P = 0.05$ (Figure 5). In addition, note that C_{lab} (based on ρ) is likely to be a considerable *under*estimate of the true rate of exchange because it ignores the added contribution of gene conversion to C. Over small physical distances (*i.e.*, on the order of ~1 kb), gene conversion is expected to contribute as much to the recombination rate as ρ (Andolfatto and Nordborg 1998).

It is difficult to imagine that recombination rates are different enough in natural populations (relative to lab strains) to account for the unusual haplotype structure observed at the breakpoint. A study of 2nd chromosome recombination rates in the F_1 progeny of lab strains and wild-caught *D. melanogaster* lines suggests that genetic background does not have a large effect (Brooks and Marks 1986). In addition, age and temperature also appear to have relatively minor effects (Ashburner 1989). Even if it is assumed that the true recombination rate is larger than C_{con}, it could be argued that local variation in mutation or recombination rates make it easier to observe the strong haplotype structure at the *In(2L)t* breakpoint. We have no reason to believe that recombination is suppressed to this degree near the breakpoint between *standard* chromosomes.

Population genetic models: There are several alternatives to the panmictic neutral model that make it more likely to observe fewer-than-expected haplotypes in a sample. These include some forms of selection and population subdivision. Epistatic selection (Lewontin 1974) or balancing selection (Strobeck 1983) can result in a deficiency in the number of haplotypes relative to neutral expectations. Other possibilities include transient selection (Hudson *et al.* 1994, 1997) or "traffic" models (Kirby and Stephan 1996).

The findings of strong associations among polymorphic sites in standard chromosomes and the extreme divergence between the two major haplotype classes (Figure 4B) raise the possibility that some form of selection on standard alleles predated the appearance of *In(2L)t*, and that these standard haplotype classes are quite old. Since we do not have a reasonable estimate of the divergence between species across this region, a plausible alternative explanation for the high level of nucleotide polymorphism at the breakpoint is that either the mutation rate or the level of selective constraint varies across the sequenced region. The possibility of a "hotspot" for mutation at the breakpoint seems unlikely given that polymorphisms in the *In(2L)t* class of chromosomes do not cluster near the location of this elevated window of polymorphism in standard chromosomes.

The finding of at least one putative exon in the sequenced region does suggest that the level of selective constraint is likely to vary. Note, however, that while heterogeneity in selective constraint may be a sufficient explanation for the peak of elevated polymorphism, it does not explain the deficiency in the number of haplotypes.

An alternative to selection models for the pattern at the *In(2L)t* breakpoint are demographic models (*cf.* David and Capy 1988). If a single population sample actually consists of individuals from two diverged subpopulations, mutations that contribute to divergence between the populations will appear to be in complete linkage disequilibrium. Strong linkage disequilibrium will result in fewer distinct haplotypes, given the number of mutations observed, than expected in a panmictic population. Recent data on the *vermilion* locus suggest a deficiency of haplotypes in non-African relative to African populations (Begun and Aquadro 1995). One possibility (as suggested by the authors) is that the African populations are closer to equilibrium while non-African populations have experienced a recent founder event. While consistent with patterns observed at other loci on the X chromosome (Begun and Aquadro 1993), this hypothesis was not supported by a recent multilocus microsatellite study (Irvin *et al.* 1998).

The main difference between selective and demographic models is that the latter are expected to affect the whole genome with equal strength, while with selection, recombination will tend to uncouple the histories of neutral sites from that of the site under selection. Thus, depending on the rate of recombination, selection models can lead to considerable heterogeneity in linkage disequilibrium across a given region (Hudson and Kaplan 1988; Hudson *et al.* 1997). The pattern in our data appears to be localized to the breakpoint. Thus, while demographic models cannot be excluded, they seem unlikely given the number of inferred recombination events and rare polymorphisms observed further from the breakpoint (see Figure 2). A similar pattern is observed at the *Sod* locus (Hudson *et al.* 1997). However, in the presence of intermediate levels of recombination, closely linked regions may show different patterns of variability even when selection is not operating. So further work is necessary to evaluate the likelihood of these (and other) data under various demographic models. Current work attempts to determine whether or not the unusual haplotype structure we detected at the breakpoint of *In(2L)t* is a general feature of *D. melanogaster* populations. Sampling standard alleles from other populations, particularly those in Africa (presumably ancestral, *cf.* Lachaise *et al.* 1988), may establish whether epistatic selection or a recent founder event are likely explanations for our data.

Implications for clines and linkage disequilibrium between *Adh* and *In(2L)t*: The Fast/Slow allozyme polymorphism of the *Adh* locus in *D. melanogaster* is in complete linkage disequilibrium with *In(2L)t* in many populations (reviewed in Lemeunier and Aulard 1992). Despite this association, Bénassi *et al.* (1993) demonstrated that among 19 restriction site variants within the *Adh* locus, only the Fast/Slow site and the Δ1 insertion-deletion were in significant linkage disequilibrium with *In(2L)t*. Thus, while all *In(2L)t* chromosomes have the *Adh-Slow* allele, there is evidence for considerable between-arrangement exchange at *Adh* (*i.e.*, shared polymorphisms; see Table 2). In fact, laboratory estimates of recombination between *Adh* and *In(2L)t* are very high ($\sim 10^{-3}$ to 10^{-4}; Malpica *et al.* 1984). Therefore, it is curious that strong linkage disequilibrium is observed between the *Adh* amino acid polymorphism and *In(2L)t*. One hypothesis is that some form of epistatic selection maintains linkage disequilibrium (Van Delden and Kamping 1989). An alternative is that *In(2L)t* and *Adh-F* alleles arose in isolation and have only recently encountered one another (Veuille *et al.* 1998).

Standard-*Adh-Fast* and *In(2L)t-Adh-Slow* alleles coexist at intermediate frequencies in many populations worldwide (Voelker *et al.* 1978; Knibb 1986; Bénassi *et al.* 1993; Veuille *et al.* 1998; S.-C. Tsaur, unpublished results). Our results suggest that while *In(2L)t* is relatively young, its age (∼1.6 million generations) is at least ∼100-fold that of the expected half-life of linkage disequilibrium between *Adh* and the inversion. Our samples of *In(2L)t* alleles (including an African population) do not differ significantly at the nucleotide level. There is therefore no evidence that *In(2L)t* remained isolated for any lengthy period of time. In the absence of selection maintaining linkage disequilibrium, demographic models would have to assume that Standard-*Adh-Fast* and *In(2L)t-Adh-Slow* arose in strongly isolated populations and have coexisted for fewer than several thousand generations.

We thank J. Comeron, R. Hudson, A. Llopart, and M. Przeworski for computer programs and helpful discussions. Comments from G. Churchill, A. Crawford, B. Charlesworth, D. Charlesworth, M. Jensen, J. Karubian, and especially M. Przeworski significantly improved this manuscript. S.-C. Tsaur kindly provided unpublished *Adh* data. We are grateful to J. Gladstone for expert technical assistance. This research was supported by National Science Foundation grant DEB-9408869 to M.K. P.A. holds a Postgraduate Scholarship from the National Science and Engineering Research Council of Canada.

LITERATURE CITED

Aguadé, M., 1988 Restriction map variation at the *Adh* locus of *Drosophila melanogaster* in inverted and noninverted chromosomes. Genetics **119:** 135–140.

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

Aquadro, C. F., S. F. Deese, M. M. Bland, C. H. Langley and C. C. Laurie-Ahlberg, 1986 Molecular population genetics of the *Alcohol dehydrogenase* gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

Aquadro, C. F., A. L. Weaver, S. W. Schaeffer and W. W. Anderson, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*—the *amylase* gene region. Proc. Natl. Acad. Sci. USA **88:** 305–309.

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

Babcock, C. S., and W. W. Anderson, 1996 Molecular evolution of the sex ratio inversion complex in *Drosophila pseudoobscura*: analysis of the *Esterase-5* region. Mol. Biol. Evol. **13:** 297–308.

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

Begun, D. J., and C. F. Aquadro, 1995 Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *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 of 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.

Brooks, L. D., and R. W. Marks, 1986 The organization of genetic variation for recombination in *Drosophila melanogaster*. Genetics **114:** 525–547.

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

Cirera, S., and M. Aguadé, 1997 Evolutionary history of the sex-peptide (*Acp70A*) gene region in *Drosophila melanogaster*. Genetics **147:** 189–197.

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.

Dinocera, P. P., and Y. Sakaki, 1990 *LINEs*: a superfamily of retrotransposable ubiquitous DNA elements. Trends Genet. **6:** 29–30.

Dobzhansky, T., 1970 *Genetics of the Evolutionary Process*. Columbia University Press, New York.

Fleischer, R. C., C. E. McIntosh and C. L. Tarr, 1998 Evolution on a volcanic conveyer belt: using phylogeographic reconstructions and K-Ar-based ages of the Hawaiian Islands to estimate molecular evolutionary rates. Mol. Ecol. **7:** 533–545.

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

Grell, R., 1962 A new model for secondary nondisjunction: the role of distributive pairing. Genetics **47:** 1737–1754.

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

Hasson, E., and W. F. Eanes, 1996 Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics **144:** 1565–1575.

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

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

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 Sci. Soc., 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, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

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, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

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.

Ishii, K., and B. Charlesworth, 1977 Associations between allozyme loci and gene arrangements due to hitch-hiking effects of new inversions. Genet. Res. **30:** 93–106.

Karlin, S., and C. Macken, 1991 Some statistical problems in the assessment of inhomogeneities in DNA sequence data. JASA **86:** 27–35.

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.

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

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.

Knibb, W. R., 1986 Temporal variation of *Drosophila melanogaster Adh* allele frequencies, inversion frequencies and population sizes. Genetica **71:** 175–190.

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.

Krimbas, C. B., and J. R. Powell, 1992 *Drosophila Inversion Polymorphism*. CRC Press, Boca Raton, FL.

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.

Leicht, B. G., E. M. S. Lyckegaard, C. M. Benedict and A. G. Clark, 1993 Conservation of alternative splicing and genomic organization of the myosin alkali light-chain (*Mlc1*) gene among Drosophila species. Mol. Biol. Evol. **10:** 769–790.

Lemeunier, F., and S. Aulard, 1992 Inversion polymorphism in *Drosophila melanogaster*, pp. 339–405 in *Drosophila Inversion Polymorphism*, edited by C. B. Krimbas and J. R. Powell. CRC Press, Boca Raton, FL.

Lewontin, R. C., 1974 The genetic basis of evolutionary change. Columbia University Press, New York.

Lyttle, T. W., and D. S. Haymer, 1992 The role of transposable element *hobo* in the origin of endemic inversions in wild populations of *Drosophila melanogaster*. Genetica **86:** 113–126.

Malpica, J. M., J. M. Vassallo, A. Frias and F. Fuentes-Bol, 1984 On recombination among *In(2L)t*, α*Gpdh* and *Adh* in *Drosophila melanogaster*. Genetics **115:** 141–142.

Mathiopoulos, K. D., A. Della Torre, V. Predazzi, V. Petrarca and M. Coluzzi, 1998 Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. Proc. Natl. Acad. Sci. USA **95:** 12444–12449.

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

Mettler, L. E., R. A. Voelkner and T. Mukai, 1977 Inversion clines in natural populations of *Drosophila melanogaster*. Genetics **87:** 169–176.

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

Mukai T., and R. A. Voelker, 1977 The genetic structure of natural populations of *Drosophila melanogaster* XIII. Further studies in linkage disequilibrium. Genetics **86:** 175–185.

Navarro, A., E. Betrán, A. Barbadilla and A. Ruiz, 1997 Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics **146:** 695–709.

Nei, M., and W. H. Li, 1980 Non-random association between electromorphs and inversion chromosomes in finite populations. Genet. Res. **35:** 65–83.

Novitski, E., and G. Braver, 1954 An analysis of crossing-over within a heterozygous inversion in *Drosophila melanogaster*. Genetics **39:** 197–209.

Ochman, H., F. J. Ayala and D. L. Hartl, 1993 Use of polymerase chain reaction to amplify segments outside boundaries of known sequences. Meth. Enzymol. **218:** 309–321.

Rozas, J., and M. Aguadé, 1994 Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of Drosophila. Proc. Natl. Acad. Sci. USA **91:** 11517–11521.

Rozas, J., C. Segarra, G. Ribo and M. Aguadé, 1999 Molecular population genetics of the *rp49* gene region in different chromosomal inversions of Drosophila subobscura. Genetics **151:** 189–202.

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

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

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal rearrangement. Genetics **103:** 545–555.

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

Swofford, D. L., 1993 *PAUP: Phylogenetic Analysis Using Parsimony*. v.3.1.1, Illinois Natural History Survey, Champaign, IL.

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.

Takano, T. S., 1998 Rate variation of DNA sequence evolution in the Drosophila lineages. Genetics **149:** 959–970.

Van Delden, W., and A. Kamping, 1989 The relation between the polymorphisms at the *Adh* and α*Gpdh* loci and the *In(2L)t* inversion in *Drosophila melanogaster* in relation to temperature. Evolution **43:** 775–793.

Van Delden, W., and A. Kamping, 1991 Changes in relative fitness with temperature among second chromosome arrangements in *Drosophila melanogaster*. Genetics **127:** 507–514.

Veuille, M., V. Bénassi, S. Aulard and F. Depaulis, 1998 Allele-specific population structure of *Drosophila melanogaster* Alcohol dehydrogenase at the molecular level. Genetics **149:** 971–981.

Voelker, R. A., C. C. Cockerham, F. M. Johnson, H. E. Schaffer, T. Mukai *et al.*, 1978 Inversions fail to account for allozyme clines. Genetics **88:** 515–527.

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.

Wesley, C. S., and W. F. Eanes, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91:** 3132–3136.

Communicating editor: G. A. Churchill