Haplotype Test Reveals Departure From Neutrality in a Segment of the *white* Gene of *Drosophila melanogaster*

David A. Kirby and Wolfgang Stephan

Department of Zoology, University of Maryland, College Park, Maryland 20742

Manuscript received April 19, 1995
Accepted for publication September 9, 1995

ABSTRACT

Restriction map studies previously revealed extensive linkage disequilibria in the transcriptional unit of the *white* locus in natural *Drosophila melanogaster* populations. To understand the causes of these disequilibria, we sequenced a 4722-bp region of the *white* gene from 15 lines of *D. melanogaster* and 1 line of *Drosophila simulans*. Statistical tests applied to the entire 4722-bp region do not reject neutrality. In contrast, a test for high-frequency haplotypes ("Haplotype test") revealed an 834-bp segment, encompassing the 3′ end of intron 1 to the 3′ end of intron 2, in which the structure of variation deviates significantly from the predictions of a neutral equilibrium model. The variants in this 834-bp segment segregate as single haplotype blocks. We propose that these unusually large haplotype blocks are due to positive selection on polymorphisms within the *white* gene, including a replacement polymorphism, Arg → Leu, within this segment.

THE *white* gene of Drosophila has played an important role in the history of genetics (see reviews by JUDD 1987; HAZELRIGG 1987). It was the first gene for which a mutation was reported in *Drosophila melanogaster* (MORGAN 1910; BRIDGES and MORGAN 1923). The gene is located on the *X* chromosome at map position 1.5 (LINDSLEY and ZIMM 1992) and in band 3C2 of the polytene chromosome map (BRIDGES 1938; LEFEVRE and WILKINS 1966). The complete sequence has been determined, encompassing 14 kb (O'HARE *et al.* 1984), and the positions of exon/intron splice junctions are known (O'HARE *et al.* 1984; PEPLING and MOUNT 1990). Sequencing of a *white* cDNA revealed that the gene product consists of 687 amino acids (PEPLING and MOUNT 1990). *white* mutants lack pigment in the compound eye, ocelli, testes sheath and Malphigian tubules. Despite the abundance of information on this gene, the functional nature of the *white* gene has not been established, although involvement in *trans*-membrane pigment precursor transport has been suggested (SULLIVAN and SULLIVAN 1975; MOUNT 1987; DREESEN *et al.* 1988; TEARLE *et al.* 1989).

A restriction map study of the *white* gene region for 64 lines from three natural populations of *D. melanogaster* by MIYASHITA and LANGLEY (1988) revealed extensive nonrandom associations (linkage disequilibria) between molecular polymorphisms throughout the transcriptional unit. These results were confirmed in a more recent restriction map survey of 200 lines from four additional natural populations (MIYASHITA *et al.* 1993). The most interesting aspects of their studies are that linkage disequilibria are clustered in a region of ~4 kb, encompassing most of intron 1 to intron 4, and that a large proportion of two-locus combinations show significant linkage disequilibrium with similar degree and direction between subpopulations. This may suggest that these disequilibria are maintained by natural selection (LEWONTIN 1974). Indeed, based on homogeneity tests of linkage disequilibria and OHTA's (1982) approach, MIYASHITA *et al.* (1993) hypothesized that epistatic selection was the cause of most linkage disequilibria. SCHAEFFER and MILLER (1993) also found clustering of linkage disequilibria, over shorter genetic distances, in the *Adh* locus of *Drosophila pseudoobscura*, which they attributed to epistatic selection.

There are a number of positive selective forces that can lead to linkage disequilibrium between polymorphisms: directional selection, balancing selection and epistatic selection (KIMURA 1956; FELSENSTEIN 1965; LEWONTIN 1974). To clarify the role of natural selection in shaping the organization of polymorphisms in the *white* gene, we examined DNA sequence variation in a 4722-bp region of the *white* gene in 15 lines of *D. melanogaster* from Beltsville, Maryland. A statistical test of neutrality developed by HUDSON *et al.* (1994) identified an 834-bp region, encompassing the 3′ end of intron 1 to the 3′ end of intron 2, in which the pattern of variation deviates from the predictions of a neutral equilibrium model. In this region, the polymorphisms form distinct high-frequency haplotypes. Several selection models that may explain these observations are discussed.

MATERIALS AND METHODS

Drosophila lines and DNA preparation: Fifteen isofemale lines of *D. melanogaster* (Bv2, Bv3, Bv4, Bv6, Bv7, Bv8, Bv10,

Corresponding author: David A. Kirby, Department of Zoology, University of Maryland, College Park, MD 20742.
E-mail: dkirby@zool.umd.edu

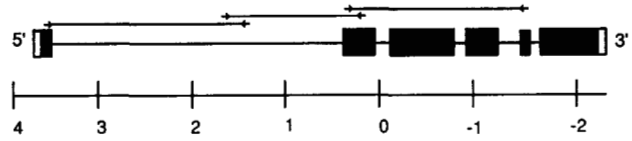


FIGURE 1.—Structure of the transcriptional unit of the *white* gene of *D. melanogaster*. The boxes represent exons. Black portions of the boxes represent coding regions, and open portions represent the untranslated regions of the gene. Introns are shown as the lines connecting the boxes. The coordinates adopted from O'HARE *et al.* (1984) are distances in kilobases. The arrows show relative positions of PCR primers. The lines connecting the PCR primers show the PCR amplification products. The portion of the gene from the beginning of the first intron to the end of the fourth intron was sequenced for each of the 15 *D. melanogaster* lines and *D. simulans*.

Bv12, Bv14, Bv19, Bv22, Bv36, Bv37, Bv38, and Bv39) and one *D. simulans* isofemale line (sim) were established from a population collected in October 1990 in Beltsville, Maryland. For *D. melanogaster*, homozygous lines for the *X* chromosome were created using balancer stock Fm7a (MERRIAM 1968) from the National Drosophila Species Resource Center (Bowling Green State University, Bowling Green, OH). For *D. simulans*, the *X* chromosome was isolated by means of an attached-X stock kindly provided by J. COYNE. Genomic DNA was purified using CsCl-Sarkosyl gradients (BINGHAM *et al.* 1981).

PCR amplification and direct DNA sequencing: For each of the 15 *D. melanogaster* lines and for *D. simulans*, three overlapping segments were amplified by PCR (SAIKI *et al.* 1988) totaling 4722 bp (Figure 1). Primers were used to initially amplify each segment, and a third internal primer was used for subsequent reamplification along with one of the initial primers. The primers for the three segments, corresponding to coordinates from the previously published *D. melanogaster white* sequence (O'HARE *et al.* 1984), are as follows (internal primers are in parentheses): 3496–3477 (3474–3455) and 1860–1841, 1960–1941 and 130–111 (150–131), and 298–279 and −1317 to −1336 (−1287 to −1306). The nucleotide sequence was determined for both strands of DNA using a set of 23 oligonucleotide primers spaced approximately every 225 bp, a set of 23 oligonucleotide primers that were the reverse complement of the previous set, and the PCR primers. These same primers were used to sequence *D. simulans*, with the addition of three primers that had to be designed from the *D. simulans* sequence itself. PCR amplification was carried out in 50-μl reaction volumes containing 200 ng of template DNA, 5 μl of Promega Mg-free 10× buffer, 4 μl of 25 mM $MgCl_2$, 50 pmol of each primer and 2 units of Promega *taq* polymerase. Thirty-five cycles were performed according to the following profile: 94°, 1 min; 55°, 2 min; 72°, 3 min. Products were reamplified under a similar profile and reaction mixture with two changes: 0.5 μl of the initial PCR product was used as template and a 65° annealing temperature was used instead of 55°.

After reamplification, double-stranded products were cleaned with two phenol/chloroform extractions, precipitated with 1 volume of ethanol and 0.1 volume 3 M NaOAc and washed with 1 ml of 70% ethanol. The dried samples were resuspended in 20 μl water and concentrations of DNA were determined by comparison to a DNA ladder of known concentration. Sequencing was performed with the dsDNA Cycle Sequencing System according to manufacturer's specifications (Life Technologies, Inc., Gaithersburg, MD). To control for artifacts due to amplification by *taq* polymerase, sites that were polymorphic in only one of the sample of 15 lines were reamplified and resequenced. The overlapping fragments were assembled using the GENEJOCKEY computer program (Biosoft, Cambridge, UK).

Sequence comparisons and analyses: Sequences were aligned by eye. From the aligned sequences of 15 *D. melanogaster* lines, we estimated the parameters θ (WATTERSON 1975; NEI 1987, Equation 10.3) and nucleotide diversity π (NEI 1987, Equation 10.6), excluding insertions and deletions. Under neutral equilibrium conditions, these are estimates of $3N\mu$ for *X*-linked genes, where N is the effective population size and μ is the neutral mutation rate. The sequences were tested for departures from the neutral expectations using the HUDSON, KREITMAN, AGUADÉ (HKA) test (HUDSON *et al.* 1987), TAJIMA's (1989) test and the test of FU and LI (1993) with no outgroup.

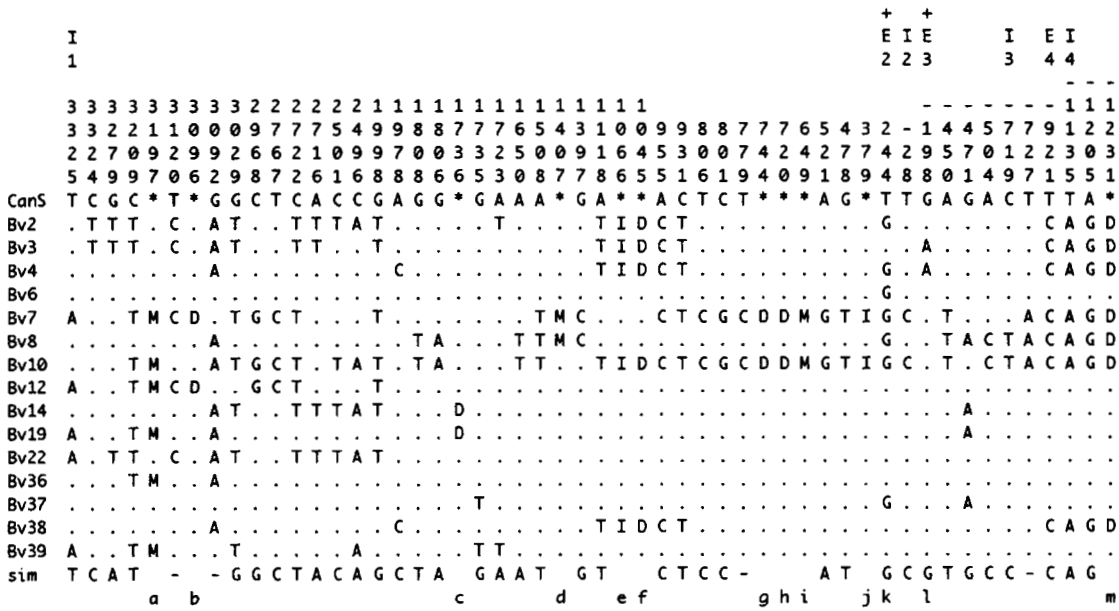
In addition, we used a test of neutrality developed by HUDSON *et al.* (1994) based on the occurrence of maximal subsets of a sample with low levels of variation, which is henceforth called the Haplotype test. The Haplotype test can estimate the probability of finding a maximal subset of a sample that has zero polymorphisms, when there are m polymorphisms in the entire sample. To test all regions of the *white* gene, the Haplotype test was applied in a sliding window across the entire sequenced portion. Windows consisted of polymorphic sites, because the only informative sites for the Haplotype test are polymorphic sites. In all of the following applications of the Haplotype test, the P values were estimated from 10,000 computer-generated samples. The computer program we used for running the Haplotype test was kindly provided by R. R. HUDSON.

To interpret the P values obtained from the Haplotype test, we attempted to generate an appropriate null distribution of the maximal subsets with zero polymorphisms (under a neutral equilibrium model such that the spatial ordering of the polymorphisms is random with regard to their frequencies). Two random datasets were produced in the following manner. First, 1000 permutations were created by shuffling the columns of the observed dataset. This creates a new set of sequences with the same base frequencies as the observed data but destroys the spatial ordering. Standard tests of neutrality have shown that the entire gene does not deviate from neutrality (see next section); therefore this procedure provides an appropriate null distribution. Second, 1000 samples were created that contained the same number of silent polymorphisms as the observed data, but the polymorphisms were randomly distributed among 15 lines using the coalescent process (HUDSON 1983). These samples were generated without recombination. This second distribution is only informative if the Haplotype test is applied as a sliding window. The Haplotype test was applied to these random datasets in the same manner it was applied to the observed data . The distribution of the lowest P values from the sliding windows of both random datasets were plotted separately for each window size.

RESULTS

DNA polymorphism at the *white* gene: Polymorphism data are summarized in Figure 2 and Table 1. There are 58 nucleotide site polymorphisms in the sample. This includes 46 nucleotide changes within the introns, 10 synonymous changes in codons (one of which is a first position change), and 2 replacement substitutions (one of which is a third position change). There are a total of 19 insertion/deletion polymorphisms. From comparisons with the *white* sequence of *D. simulans* (Figure 2), we can infer that six polymorphisms are due to deletions and eight are due to insertions; no inference

a.



b.

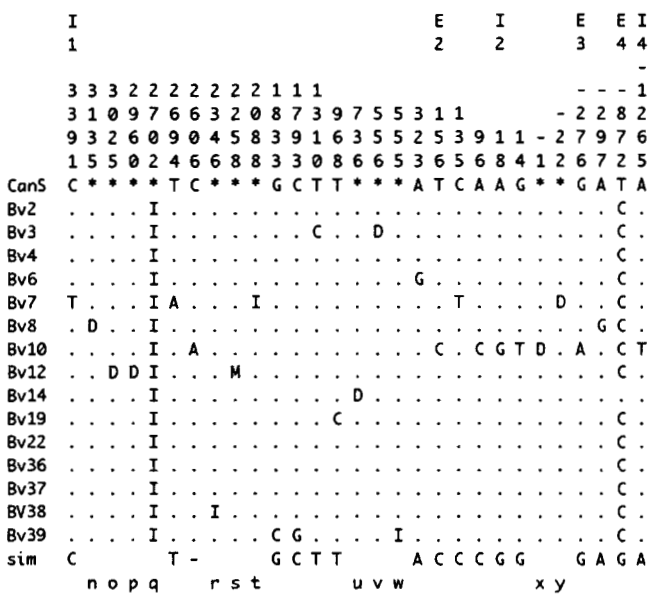


FIGURE 2.—DNA polymorphisms found in 15 lines of *D. melanogaster*. (a) Polymorphisms found in more than one line. (b) Polymorphisms found in only one line. The nucleotides within the reference sequence Canton-S (CanS) (O'HARE *et al.* 1984) are along the top. Asterices indicate insertions (I), deletions (D) or complex mutations (M). The numbers above the sequence represent the position numbers of each segregating site within the reference sequence, the nucleotide preceding an insertion, the start of a deletion or the nucleotide preceding a complex mutational event. The homologous nucleotides within the *D. simulans* sequence (sim) are along the bottom. Dashes indicate where the homologous nucleotide of *D. simulans* could not be determined. Sequence domains of the *white* gene are indicated at the top, E (exon) and I (intron). The symbol (+) indicates amino acid replacement polymorphisms at sites 244 and −198. Insertions, deletions, complex mutations and amino acid changes are as follows: a, 3197–3191 replaced by GAAAAA; b, 3096–3092; c, 1736; d, 1407–1402 replaced by TTTTTTTTTTTT; e, TAG; f, 1045–1032; g, 744–739; h, 720–714; i, 649–643 replaced by CTTTTT; j, GGGTACT; k, Arg (CGG) to Leu (CUG); l, Met (AUG) to Ile (AUA); m, −1231 to −1247; n, 3135–3131; o, 3025–3010; p, 2960–2955; q, C; r, ATT; s, 2258–2218 replaced by TCGAAAAAT; t, G; u, 736–730; v, 556; w, G; x, −1 to −2; y, −22.

about the other five can be made. Labeling of insertion/deletion within the figure is relative to the reference sequence, Canton-S (O'HARE *et al.* 1984) and not to the *D. simulans* sequence. One additional insertion was found in all 15 sequences but was not present in the reference sequence. In addition, there are four cases of

TABLE 1

Variation in the homopolymeric regions of the *white* gene

	Coordinates of homopolymeric regions		
	2218	1886	573
CanS	$(A)_{10}$	$(T)_{10}$	$(G)_{10}T(G)_5$
Bv2	*	*	*
Bv3	*	*	$(G)_{10}T(G)_6$
Bv4	*	*	$(G)_{12}T(G)_5$
Bv6	*	*	$(G)_{10}T(G)_6$
Bv7	$(A)_{12}$	*	$(G)_{13}$
Bv8	*	*	*
Bv10	*	$(T)_9$	$(G)_{13}$
Bv12	*	$(T)_{11}$	$(G)T(G)_6$
Bv14	*	*	$(G)_8T(G)_5$
Bv19	*	*	$(G)_{13}$
Bv22	*	$(T)_9$	$(G)_9T(G)_5$
Bv36	*	*	$(G)_3A(G)_{12}$
Bv37	*	*	$(G)_{12}T(G)_5$
Bv38	*	*	$(G)_{12}T(G)_5$
Bv39	*	*	*

Coordinates (from O'HARE *et al.* 1984) represent the first nucleotide of the homopolymeric region. The repeated base of the homopolymeric region is given in parentheses with the number of repeats as subscripts. Asterices indicate the homopolymeric region is the same as the reference sequence Canton-S (CanS).

polymorphisms involving complex mutational events. In these cases there is evidence that both an insertion/deletion event took place, as well as a nucleotide site change(s). Two of the four complex mutations involve a single base change accompanied by a single base insertion/deletion. Three areas in the region sequenced could be classified as homopolymeric regions. These regions are composed of a single nucleotide base repeated 10 or more times. Variation within the homopolymeric regions is summarized in Table 1.

Estimates of Π and θ for the 15 *D. melanogaster* sequences are as follows: $\hat{\Pi} = 0.0046$; $\hat{\theta} = 0.0048$. These estimates are based on the number of silent sites, 3920 bp, and the number of silent segregating sites, 61. The number of silent segregating sites includes two polymorphic sites within the homopolymeric regions, and three of the four complex mutations because there is a single nucleotide site change involved in these events. The number of silent sites includes the number of bases in the introns, fourfold and twofold degenerate third position sites, twofold degenerate first position sites and excludes third positions of the codon AUG and UGG.

We found roughly equal numbers of transitional and transversional changes in the sequences. However, if we examine specific types of transitions and transversions, there appears to be an excess of A-T changes in the sample. We tested this by calculating the expected number of A-T changes using the inferred percentages of base changes from GOJOBORI *et al.* (1982) and LI *et al.* (1984). There are a total of 11 A-T changes in the sample with 61 silent changes, as opposed to an expected number of 6.16 (10.1%). This difference is significant using a chi square test with 1 d.f. ($\chi^2 = 3.85$; $P < 0.05$). We redid the test ignoring silent changes within codons, which may have constraints due to codon bias, which resulted in a larger chi square value (51 total changes with 10 A-T changes, 5.15 expected; $\chi^2 = 5.31$; $P < 0.05$).

Neutrality tests: We applied the standard neutrality tests to the entire 4722-bp region. TAJIMA's (1989) test, which compares the number of pairwise differences to the number of segregating sites, failed to reject the null hypothesis of neutrality ($D = -0.18$, NS). Results of the FU and LI (1993) test also failed to reject neutrality ($D^* = 0.09$, NS). The HKA test (HUDSON *et al.* 1987) can also be used to detect deviations from neutrality by comparing levels of divergence and levels of polymorphism for two different loci. We applied this test using the 5′ flanking region of *Adh* as the second locus. Equations 5 and 6 from Hudson *et al.* (1987) must be modified if the HKA test is used to compare autosomal and *X*-linked genes (BEGUN and AQUADRO 1991). Applying the modified HKA test results in a chi square value of 0.02 (1 d.f.) that is nonsignificant. Although neutrality could not be rejected using these standard statistical tests over the entire 4722 bp, there are segments within the *white* gene that deviate significantly from neutrality (as described below).

Recombination and the structure of variation in the *white* gene: Recombination allows different segments of a gene to have different evolutionary histories (HUDSON 1983). Therefore, certain selection events that occur within genes of high or moderate recombination rates may not be detected by standard neutrality tests. The method of HUDSON and KAPLAN (1985) was used to infer that there were a minimum of 13 recombination events in the history of our sample. These 13 inferred recombination events occurred in the intervals 3325–3279, 3209–3120, 3120–3092, 3092–3029, 3029–2968, 2716–2501, 2496–1978, 1723–1508, 1508–1118, 931–244, 244 to −198, −450 to −471 and −471 to −504 where the numbers refer to the nucleotide site positions as in Figure 2. The method of HUDSON (1987) can be used to estimate $C = 3Nc$ for *X*-linked genes, where c is the frequency of recombination between adjacent nucleotides. When we applied this method to the *white* gene data, not including an 834-bp region that may be under the influence of selection (discussed below), we obtained an estimate of 33.58. The ratio of C/θ is estimated to be 1.78, implying that recombination is approximately twice as frequent as mutation on a per nucleotide basis. With the 834-bp region, C was estimated as 16.15. The *white* gene is at cytological position 3C2. For this region, an adjusted coefficient of exchange (ACE × 100) of 6.17 was estimated (E. C. KINDAHL and C. F. AQUADRO, personal communication). These results indicate that the recombination rate of *white* is comparable with other genes of *D. melanogaster* with moderate rates of recombination. As a comparison,

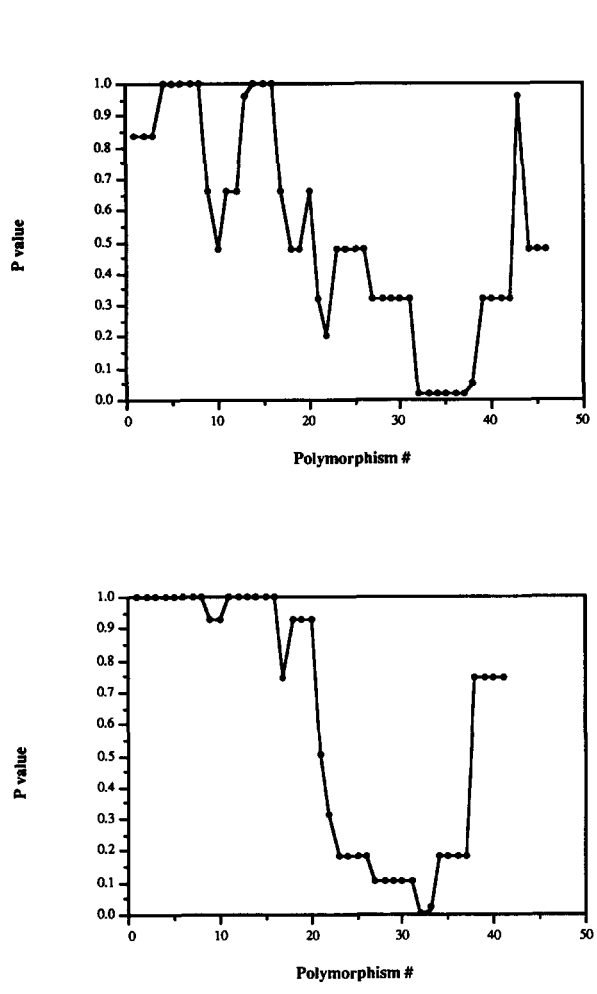


FIGURE 3.—*P* values for sliding windows of the Haplotype test across the sequenced portion of the *white* gene. The nonunique silent polymorphisms are numbered along the *x*-axis. Polymorphism number 32 corresponds to the polymorphism at coordinate 806. (a) Window size of five polymorphisms. (b) Window size of 10 polymorphisms.

in the *Adh* region the value of ACE (× 100) is 4.32, and the C/θ ratio was estimated by HUDSON (1987) to be 1.6.

We tested the neutral equilibrium hypothesis using the Haplotype test developed by HUDSON *et al.* (1994). We applied the Haplotype test to the sequenced portion of the *white* gene for all possible window sizes (2–50) of nonunique polymorphisms without recombination. There were exactly seven window sizes (5–11) that contained regions with *P* values lower than 5% of both randomly generated datasets. Figure 3 shows the results for two window sizes. For all seven window sizes, an 834-bp segment between coordinates 806 and −28 was the only region to have *P* values 5% of both randomly generated datasets. This 834-bp segment contains 10 nonunique polymorphisms that appear to be segregating as a single haplotype block. Including unique polymorphisms and recombination in the analysis did not change these results.

The overall lowest *P* value was obtained with a window size of 10 polymorphisms, which corresponds to the entire 834-bp segment between coordinates 806 and −28. In fact, with a window size of 10 polymorphisms, none of the randomly generated samples from either dataset had any regions with lower *P* values than the observed *P* value for the 834-bp region. In this region, the sample for the nonunique silent polymorphisms consisted of 15 sequences with 10 polymorphic sites. There was a maximal subset of this sample consisting of 13 sequences with no polymorphisms. With no recombination, the estimated *P* value was 0.003. If we include unique polymorphisms, our sample consisted of 15 sequences with 21 polymorphic sites. There was a maximal subset of this sample consisting of nine sequences with no polymorphisms. With no recombination, the estimated *P* value was 0.02. With recombination, the *P* values for this region were lower. Our estimate of $3Nc$ per nucleotide for the *white* gene using the method of HUDSON (1987) was 0.008. With this recombination value, the *P* values were 0.0028 and 0.018, respectively. The results of the Haplotype tests indicate that the lack of variation in a subset of our sample between coordinates 806 and −28 is due to forces other than mutation and drift alone. We must, therefore, explore alternative hypotheses to explain the high frequency of some haplotypes within this segment.

DISCUSSION

We observed within the *white* gene an 834-bp segment between coordinates 806 and −28 that has 10 nonunique silent polymorphisms that appear to be segregating as a single haplotype. In contrast to the rest of the sequenced portion of the gene, the results of a statistical test for high-frequency haplotypes (the Haplotype test) developed by HUDSON *et al.* (1994) indicate that the haplotypes of high frequency within this 834-bp segment cannot be explained by a neutral equilibrium model of mutation and drift. The nonunique polymorphisms reveal three distinct haplotypes in this segment, as follows: (1) identical to the reference sequence Canton-S (O'HARE *et al.* 1984), frequency 8/15; (2) differs from Canton-S at one site, which results in a replacement substitution at amino acid residue 49 (PEPLING and MOUNT 1990), frequency 5/15; and (3) differs from Canton-S by 11 sites, which includes the replacement substitution, frequency 2/15. We designate these haplotypes 244T (Bv3, Bv12, Bv14, Bv19, Bv22, Bv36, Bv38, and Bv39), 244G (Bv2, Bv4, Bv6, Bv8, and Bv37) and 244G11 (Bv7and Bv10), respectively.

How might one account for the presence of high-frequency haplotypes within an 834-bp segment of the *white* gene? First, we consider some neutral hypotheses. The null hypothesis for the Haplotype test is a neutral equilibrium model in a panmictic population without geographic subdivision. STROBECK (1987) has shown that the number of haplotypes within subpopulations will tend to be lower for a given level of nucleotide

heterozygosity, compared with panmictic models. In addition, OHTA (1982) has demonstrated that limited migration between subdivided populations may cause linkage disequilibrium. It is possible to determine whether limited migration is the cause of linkage disequilibrium by using a variance of linkage disequilibrium analysis developed by OHTA (1982). MIYASHTA *et al.* (1993) used OHTA's (1982) analysis to show that limited migration is not the cause of most significant linkage disequilibria they found within the transcriptional unit of the *white* gene. We identified DNA polymorphisms in the 834-bp region corresponding to most of the restriction polymorphisms found by MIYASHITA and LANGLEY (1988) and MIYASHITA *et al.* (1993)

The pattern of high-frequency haplotypes in our data might also be explained by a neutral model with a history of bottlenecks and/or population expansions. TACHIDA (1994) pointed out that linkage disequilibria can be maintained for long periods of time if a population has a history of size fluctuations. Although it is difficult to examine population size changes in the past, this explanation is unsatisfactory for our data. If population fluctuations within *D. melanogaster* have caused the high-frequency haplotypes, this same pattern should be observed throughout the entire gene. However, results using a sliding window of the Haplotype test over the entire sequenced portion of the gene (Figure 3) show that this pattern is not observed in other regions of the *white* gene.

We now consider some alternative hypotheses involving selection. We only discuss hypotheses involving selection within the 834-bp segment where we observed the high-frequency haplotypes. It is possible that selection acting on variants outside the region sequenced may result in the pattern we observe in our data. However, hypotheses involving selection outside the region sequenced must explain why the same pattern is not found over the entire sequenced portion of the *white* gene (Figure 3).

The presence of two distinct haplotypes in our sample may suggest an old balanced polymorphism. However, there is not enough variation within haplotypes 244T/244G to support this hypothesis. What form of selection, then, could account for the high-frequency haplotypes in the 834-bp segment? One hypothesis is that either directional selection or a new balanced polymorphism could explain the high-frequency haplotypes. It is not possible, however, to distinguish between these two types of selection. If it is directional selection, the observed pattern is a temporary one. In this case, in a fairly short time the sweeping variant will reach a frequency of one, and most of the variation surrounding it will be eliminated. If it is a new balanced polymorphism, the new variant has recently arisen to an intermediate frequency, where it may now be at or near its equilibrium value.

To explain the high-frequency haplotypes within the 834-bp segment by balancing or directional selection we suggest the following scenario, which is similar to the scenario that HUDSON *et al.* (1994) proposed for the *Sod* locus. From comparison with the *D. simulans* sequence (Figure 2), we infer that amino acid residue 49 in haplotype 244G (Arg) may be the ancestral amino acid, although it is possible that this variant is also polymorphic within *D. simulans*. Haplotype 244G rose in frequency due to recent positive selection on a site(s) in intron 1 or intron 2. Then a mutation occurred at coordinate 244 from G → T, resulting in an Arg → Leu replacement. Subsequently, either because of selection on the amino acid variant itself or simply by hitchhiking associated with selection on 244G, the frequency of haplotype 244T also rose in frequency. This scenario explains both the high frequency of haplotype 244G and the high frequency of haplotype 244T. KAPLAN *et al.* (1989) showed that a mutant with selective advantage of 0.001 that sweeps through a population can reduce variation up to 1000 bp away from the site, assuming typical levels of *D. melanogaster* recombination. This scenario implies selection on silent changes within introns. Although we have no evidence for such mechanisms within the *white* gene, many large Drosophila introns have been shown to contain enhancer elements (BINGHAM *et al.* 1988).

A second hypothesis is that the high-frequency haplotypes are maintained by epistatic fitness interactions between polymorphisms. Epistatic interactions are expected to lead to nonrandom associations between polymorphisms and thus possibly to haplotype blocks (KIMURA 1956; LEWONTIN 1974). Ignoring the replacement substitution, there are two haplotypes for the nonunique silent polymorphisms in the segment between coordinates 806 and −28: one haplotype that is identical to Canton-S (244G or 244T) and a second haplotype that contains all 10 nonunique silent polymorphisms (244G11). If there are epistatic fitness interactions between polymorphic sites, haplotypes that are formed due to recombination between two selected haplotypes are likely to be eliminated. Therefore, recombinant haplotypes are expected to be in low frequency in the population.

There are two criteria that may be used to distinguish between the two selection hypotheses: the amount of recombination between the haplotypes and the amount of nucleotide diversity within each haplotype relative to the average nucleotide diversity within the 834-bp region. The epistatic selection hypothesis predicts that recombinants between haplotypes 244G/244T and 244G11 are likely to be eliminated and thus would be in low frequency, in contrast to the directional/balancing selection hypothesis. The selected events postulated in the directional/balancing selection hypothesis would have been very recent. Therefore, one would expect little variation in either haplotype 244G or 244T, as they have risen to high frequency recently. Haplotype 244G11 would be an older haplotype and thus should be more variable. The epistatic selection hypothesis pre-

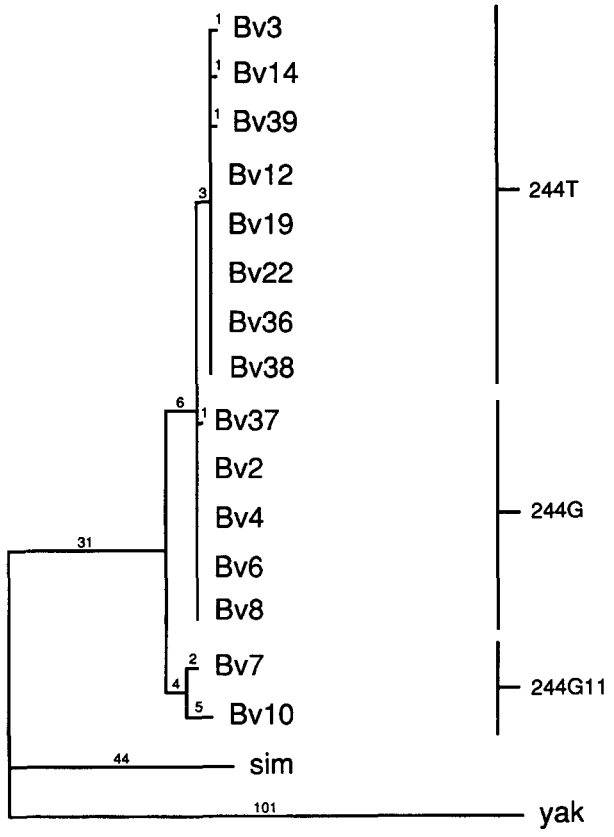


FIGURE 4.—Phylogram of the single most parsimonious tree for the 834-bp region between coordinates 806 and −28. In this analysis, insertions/deletion and complex mutational events were treated as single characters and amino acid replacements were weighted 3:1 against silent changes. *D. yakuba* (yak; D. A. KIRBY and W. STEPHAN, unpublished data) was used as an outgroup to root the tree. Numbers adjacent to each branch are the numbers of inferred mutational steps. The three haplotypes are indicated on the right.

dicts low variation within haplotypes 244G/244T and also within haplotype 244G11 compared with the average level of variation within the 834-bp region.

To examine these two criteria, maximum parsimony analysis was carried out on the 834-bp region for all 15 *D. melanogaster* lines and *D. simulans* using PAUP (SWOFFORD 1991) with the branch-and-bound option (HENDY and PENNY 1982). Exactly one most parsimonious tree was found with a length of 22 steps (discounting weighting values) for the *D. melanogaster* lines (Figure 4), which is equal to the number of mutations observed in the data. This suggests that neither recombination nor multiple mutations must be invoked to explain the data (HILTON *et al.* 1994; LEICHT *et al.* 1995). Haplotypes 244T/244G and 244G11 form two distinct clades. There are 10 steps between the two clades. In contrast, there are five steps (discounting weighting values) within the 244T/244G clade and seven steps within the 244G11 clade. This suggests that variation within each haplotype is lower than the average variation of the 834-bp region, which supports the epistatic selection hypothesis. This observation, in addition to the lack of any observed recombinants within the 834-bp region, provides support for the epistatic selection hypothesis. It should be pointed out, however, that variation within haplotypes 244G (one step) and 244T (four steps) is lower than within haplotype 244G11 (seven steps), which is consistent with the directional/balancing selection hypothesis. Although our data provide more support for the epistatic selection hypothesis, a larger sample size is needed to distinguish between these hypotheses.

If epistatic selection is responsible for the observed pattern within the 834-bp region as our data suggest, a potential mechanism for epistatic fitness interactions could involve secondary structures of precursor mRNA (STEPHAN and KIRBY 1993; KIRBY *et al.* 1995). In its simplest form, the mechanism underlying this interaction may be as follows: a mutation occurring in a secondary structural element such as the helix of an RNA hairpin may be individually deleterious, because it increases the structure's free energy, which may destabilize this structure. However, the pairing potential of a functionally important structure and, thus, the fitness can be restored if a second "compensatory" mutation occurs in the complementary sequence of the helix. At present it is unclear whether such a mechanism is operating within the *white* gene.

We thank JOHN BRAVERMAN, BRIAN CHARLESWORTH, ANDY CLARK, JODY HEY, MANYUAN LONG, STEVE MOUNT, SPENCER MUSE, JOHN PARSCH and an anonymous reviewer for their comments and suggestions on this manuscript. We thank PAUL LEWIS for help with phylogenetic reconstruction. This research was supported by grants from the National Institutes of Health (GM-46233) and the National Science Foundation (DEB-9407226).

LITERATURE CITED

BEGUN, D. J., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the *X* chromosome in Drosophila: evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

BINGHAM, P. M., R. LEVIS and G. M. RUBIN, 1981 Cloning of DNA sequences from the *white* locus of *Drosophila melanogaster* by a novel and general method. Cell **25:** 693–704.

BINGHAM, P. M., T.-B. CHOU, I. MIMS and Z. ZACHAR, 1988 On-off regulation of gene expression at the level of splicing. Trends Genet. **4:** 134–138.

BRIDGES, C. B., 1938 A revised map of the salivary gland X chromosome of *Drosophila melanogaster.* J. Hered. **29:** 11–13.

BRIDGES, C. B., and T. H. MORGAN, 1923 *The Third-Chromosome Group of Mutant Characters of Drosophila melanogaster.* Carnegie Institute, Washington, DC. [Publication no. 327.]

DREESEN, T. D., D. H. JOHNSON and S. HENIKOFF, 1988 The brown protein of *Drosophila melanogaster* is similar to the white protein and to components of active transport complexes. Mol. Cell. Biol. **8:** 5206–5215.

FELSENSTEIN, J., 1965 The effects of linkage on directional selection. Genetics **52:** 349–363.

FU, Y.-X., and W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

GOJOBORI, T., W.-H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

HAZELRIGG, T., 1987 The *Drosophila white* gene: a molecular update. Trends Genet. **3:** 43–47.

HENDY, M. D., and D. PENNY, 1982 Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosci. **59:** 277–290.

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

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**: 283–201.

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

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116**: 153–159.

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.

JUDD, B. H., 1987 The *white* locus of *Drosophila melanogaster*, pp. 81–94 in *Results and Problems in Cell Differentiation 14. Structure and Function of Eukaryotic Chromosomes*, edited by W. HENNING. Springer-Verlag, Berlin.

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., S. V. MUSE and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92**: 9047–9051.

LEFEVRE, G., and M. D. WILKINS, 1966 Cytogenetic studies on the *white* locus in *Drosophila melanogaster*. Genetics **53**: 175–187.

LEICHT, B. G., S. V. MUSE, M. HANCZYC and A. G. CLARK, 1995 Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. Genetics **139**: 299–308.

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

LI, W.-H., C.-I. WU and C.-C. LUO, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**: 58–71.

LINDSLEY, D. L., and G. G. ZIMM, 1992 *The Genome of Drosophila melanogaster*. Academic Press, San Diego.

MERRIAM, J. R., 1968 New mutants report. Dros. Inf. Serv. **43**: 64.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120**: 199–212.

MIYASHITA, N. T., M. AGUADÉ and C. H. LANGLEY, 1993 Linkage disequilibrium in the *white* locus region of *Drosophila melanogaster*. Genet. Res. **62**: 101–109.

MORGAN, T. H., 1910 Sex limited inheritance in *Drosophila*. Science **32**: 120–122.

MOUNT, S. M., 1987 Sequence similarity. Nature **325**: 487.

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

O'HARE, K., C. MURPHY, R. LEVIS and G. M. RUBIN, 1984 DNA sequence of the *white* locus of *Drosophila melanogaster*. J. Mol. Biol. **180**: 437–455.

OHTA, T., 1982 Linkage disequilibrium due to random drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79**: 1940–1944.

PEPLING, M., and S. M. MOUNT, 1990 Sequence of a cDNA from the *Drosophila melanogaster white* gene. Nucleic Acids Res. **18**: 1633.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, *et al.* 1988 Primer-directed enzymatic amplifications of DNA with a thermostable DNA polymerase. Science **239**: 487-491.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135**: 541-552.

STEPHAN, W. and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135**: 97–103.

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

SULLIVAN, D. T., and M. C. SULLIVAN, 1975 Transport defects as the physiological basis for eye color mutants of *Drosophila melanogaster*. Biochem. Genet. **13**: 603- 613.

SWOFFORD, D. L., 1991 PAUP: phylogenetic analysis using parsimony, version 3.1. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

TACHIDA, H., 1994 Decay of linkage disequilibrium in a finite island model. Genet. Res. **64**: 137–144.

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

TEARLE, R. G., J. M. BELOTE, M. MCKEOWN, B. S. BAKER and A. J. HOWELLS, 1989 Cloning and characterization of the *scarlet* gene of *Drosophila melanogaster*. Genetics **122**: 595–606.

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

Communicating editor: A. G. CLARK