Multi-Locus Selection and the Structure of Variation at the *white* Gene of *Drosophila melanogaster*

David A. Kirby and Wolfgang Stephan

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

Manuscript received February 23, 1996

Accepted for publication June 21, 1996

ABSTRACT

We surveyed sequence variation and divergence for the entire 5972-bp transcriptional unit of the *white* gene in 15 lines of *Drosophila melanogaster* and one line of *D. simulans*. We found a very high degree of haplotypic structuring for the polymorphisms in the 3' half of the gene, as opposed to the polymorphisms in the 5' half. To determine the evolutionary mechanisms responsible for this pattern, we sequenced a 1612-bp segment of the *white* gene from an additional 33 lines of *D. melanogaster* from a European and a North American population. This 1612-bp segment encompasses an 834-bp region of the *white* gene in which the polymorphisms form high frequency haplotypes that cannot be explained by a neutral equilibrium model of molecular evolution. The small number of recombinants in the 834-bp region suggests epistatic selection as the cause of the haplotypic structuring, while an investigation of nucleotide diversity supports a directional selection hypothesis. A multi-locus selection model that combines features from both hypotheses and takes the recent history of *D. melanogaster* into account may be the best explanation for these data.

TO elucidate the mechanisms of molecular evolution, it is not only necessary to know how much genetic variation exists in natural populations, but also how the variation is structured within the genome. Population genetics theory has focused on two categories of evolutionary forces that affect the structure of genetic variation. The first category of mechanisms are selectively neutral, such as random genetic drift, restricted migration among subdivided populations or reduced recombination (HILL and ROBERTSON 1968; OHTA and KIMURA 1969; OHTA 1982; STROBECK 1987; TACHIDA 1994). The second category involves various forms of natural selection, including balancing, directional, or epistatic selection (KIMURA 1956; FELSENSTEIN 1965; LEWONTIN 1974; ASMUSSEN and CLEGG 1981).

Although the clusters of linkage disequilibrium observed in some genes, such as the *Adh* gene of *Drosophila pseudoobscura* (SCHAEFFER and MILLER 1993; KIRBY *et al.* 1995), have been attributed to one major evolutionary mechanism, it is not clear what forces cause the pattern of nonrandom associations observed in several other recent surveys of DNA sequence variation, such as the *white* gene (MIYASHITA and LANGLEY 1988; MIYASHITA *et al.* 1993; KIRBY and STEPHAN 1995) and the esterase 6 locus in *D. melanogaster* (ODGERS *et al.* 1995). It is becoming evident, however, that the structure of genetic variation within a gene may be determined by more than a single evolutionary force. For example, high levels of nonrandom associations observed in non-African popu-

lations of *D. melanogaster* at the *vermillion* locus have been attributed to a combination of selection and geographic differentiation (BEGUN and AQUADRO 1995). In addition, it has been shown that multiple polymorphisms within the *Adh* region of *D. melanogaster* affect ADH activity levels, which may translate into fitness differences (LAURIE-AHLBERG 1985; LAURIE *et al.* 1991; LAURIE and STAM 1994). It has been postulated that the structure of variation at the *Adh* gene in *D. melanogaster* is determined by a combination of balancing selection and epistatic interactions (KREITMAN and HUDSON 1991; LAURIE and STAM 1994). There is also evidence that multiple selected substitutions may occur simultaneously within small genomic segments in other species with a wide geographic range, such as *D. ananassae* (STEPHAN and MITCHELL 1992).

The *white* gene of *D. melanogaster* provides a unique opportunity to examine the evolutionary forces maintaining distinct haplotype structures and linkage disequilibria, as they extend over large distances (MIYASHITA *et al.* 1993). Here we describe an analysis of sequence variation and divergence over the entire 5972-bp transcriptional unit of the *white* gene in 15 lines of *D. melanogaster* and one line of *D. simulans*. A previous study has revealed haplotypic structuring of polymorphisms in an 834-bp segment of *white* that could not be explained by a neutral equilibrium model (KIRBY and STEPHAN 1995). To elucidate the evolutionary mechanisms responsible for the high degree of haplotypic structuring, we sequenced a 1612-bp segment of the *white* gene from an additional 33 lines of *D. melanogaster* from a European and a North American population, which encompasses this 834-bp region.

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

MATERIALS AND METHODS

Drosophila lines and DNA sequencing: Nine isofemale *D. melanogaster* lines were established in October 1990 from a population in Beltsville, Maryland (Bv5, Bv9, Bv13, Bv15, Bv16, Bv23, Bv30, Bv31 and Bv32). The Fm7a balancer (MERIAM 1968) was used to make homozygous X-chromosome lines. Twenty-four lines of *D. melanogaster* from two European populations, La Rábida, Huelva, Spain (XH1, XH3, XH11, XH14, XH18, XH19, XH23, XH30, XH34, XH36, XH44, XH48, XH55, XH61, XH62, XH83, XH84 and XH118) and Groningen, Holland (XG2, XG8, XG10, XG17, XG30 and XG67), were previously described in MARTÍN-CAMPOS *et al.* (1992). Fifteen lines of *D. melanogaster* (Bv2, Bv3, Bv4, Bv6, Bv7, Bv8, Bv10, Bv12, Bv14, Bv19, Bv22, Bv36, Bv37, Bv38, Bv39) and one *D. simulans* line (sim) are the same as those previously analyzed by KIRBY and STEPHAN (1995). Genomic DNA was purified using CsCl-Sarkosyl gradients (BINGHAM *et al.* 1981).

To complete our survey of DNA variation in the transcriptional unit of the *white* gene, two nonoverlapping segments were amplified by PCR (SAIKI *et al.* 1988) for each of the 15 *D. melanogaster* lines and for the *D. simulans* line used in KIRBY and STEPHAN (1995). The primers for each segment, corresponding to coordinates from the previously published *D. melanogaster white* sequence (O'HARE *et al.* 1984), were as follows (internal primers are in parentheses): 3791–3772 (3764–3745) and 3254–3235; –1007 to –1026 (–1029 to –1048) and –2251 to –2271. The nucleotide sequence for each segment was determined for both strands of DNA using a set of 14 sense and antisense primers, and the PCR primers. These same primers were used to sequence the *white* gene in *D. simulans*, with the addition of one primer that had to be designed from the *D. simulans* sequence itself. For the other 33 lines of *D. melanogaster*, one segment of length 1612 bp, encompassing the 3' half of the first intron to the end of the second intron, was amplified and sequenced. The primers used were 1558–1536 and –33 to –52 (–63 to –82). The nucleotide sequence for each segment was determined with a set of 12 sense and antisense primers, and the PCR primers. Procedures and conditions for PCR amplification and DNA sequencing were the same as in KIRBY and STEPHAN (1995).

DNA sequence comparisons and analyses: Sequences were aligned by eye. From the aligned sequences of 15 *D. melanogaster* lines, we estimated the parameter θ (WATTERSON 1975; NEI 1987, Equation 10.3) for the complete transcriptional unit and for each domain of the *white* gene, excluding insertions and deletions. Under neutral equilibrium conditions θ is an estimate of $3N\mu$ for X-linked genes, where N is the effective population size and μ is the neutral nucleotide mutation rate. The sequences were tested for departures from selective neutrality using the HUDSON, KREITMAN and 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 the haplotype test of neutrality, which is based on the occurrence of maximal subsets of a sample with low levels of variation (HUDSON *et al.* 1994; KIRBY and STEPHAN 1995). Null distributions for sliding windows of the haplotype test were generated as in KIRBY and STEPHAN (1995).

RESULTS

Nucleotide sequence variation and divergence in the transcriptional unit of the *white* gene: Figure 1 shows polymorphism data for the entire *white* transcriptional unit. There was a total of 75 silent segregating sites, two amino acid replacement polymorphisms, 18 length

polymorphisms and five polymorphisms resulting from complex mutational events (see Figure 1 and KIRBY and STEPHAN 1995) over the entire 5972-bp transcriptional unit in the sample of 15 *white* sequences. A single base insertion (at position 2702), a nucleotide difference (position –1582) and a single nucleotide deletion (position –2199) were found in all 15 lines we surveyed and in the cDNA sequence of PEPLING and MOUNT (1990) but not in the reference sequence. Estimates of π and θ for the 15 *D. melanogaster* sequences are as follows: $\hat{\pi} = 0.0049$; $\hat{\theta} = 0.0049$. These estimates are based on the number of silent sites, 4674 bp, and the number of silent segregating sites, 75 bp. The number of silent sites includes the number of bases in the introns, fourfold and twofold degenerate third codon position sites, twofold degenerate first position sites and excludes third positions of the codons AUG and UGG. The number of silent segregating sites includes three polymorphic sites within the homopolymeric regions and three of the five polymorphisms resulting from complex mutations because there is a single nucleotide site change involved in these events. Table 1 summarizes silent site polymorphism and average nucleotide diversity (per site) for each domain of the *white* gene.

Heterogeneity of silent site polymorphism between the structural domains of *white* was examined using a goodness of fit test (KREITMAN and HUDSON 1991). The calculated chi-squared value for heterogeneity of silent site polymorphism between individual domains was nonsignificant ($\chi^2 = 14.76$; 12 d.f.; NS). Although this test failed to reveal evidence for between-domain heterogeneity, there were several domains whose estimated θ 's were much higher than the estimated θ for the entire gene (Table 1). Therefore, we grouped the structural domains into four classes: small introns, large introns, exons, and untranslated regions (UTRs). Due to possible differences in splice site/branch point recognition and associated small nuclear ribonucleoproteins (MOUNT *et al.* 1992; TALERICO and BERGET 1994) and/or differences in regulatory roles (BINGHAM *et al.* 1988), introns were grouped according to size: small (≤ 90 bp) and large (> 90 bp) (MOUNT *et al.* 1992; STEPHAN *et al.* 1994). The calculated chi square showed significant heterogeneity of silent site polymorphism between the grouped domains ($\chi^2 = 9.82$; 3 d.f.; $P < 0.025$). An examination of the individual deviations from the expected values (Table 1) revealed that the large chi-squared value can be attributed almost solely to the small introns (≤ 90 bp). We then investigated whether the estimated values for each small intron (*i.e.*, introns 2, 3, and 5) were compatible with the average nucleotide diversity of the entire gene using the method of TAVARÉ (1984). For a zero-recombination model, TAVARÉ (1984) developed an explicit expression for calculating the probability, P , of having j segregating sites in a sample of size n for a given value of average nucleotide diversity, θ (per region). We calculated P for each of

a.

[illegible]

b.

[illegible]

FIGURE 1.—DNA polymorphisms found in the *white* gene of 15 lines of *D. melanogaster*. Polymorphisms from coordinates 3325 to -1265 are from KIRBY and STEPHAN (1995). (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. An asterisk indicates a complex mutational event (M), insertion (I) or deletion (D), which are labeled as they relate to the reference sequence. The numbers above the sequence represent the position numbers of each segregating site within the reference sequence, the nucleotide preceding a complex mutational event, the start of a deletion or the nucleotide preceding an insertion. Sequence domains of the *white* gene are indicated at the top, E (exon), I (intron) and U (untranslated region). The homologous nucleotides within the *D. simulans* sequence (sim) are along the bottom; only those sites are shown that are polymorphic in *D. melanogaster*. Dashes indicate where the homologous nucleotide of *D. simulans* could not be determined. The symbol (+) indicates amino acid replacement polymorphisms. The solitary complex mutational event found in the present study involves the replacement of positions 3564 and 3465 of the reference sequence by TG, and the only deletion involved a single base pair (position -2199). There is a homopolymeric region at position -2129, where the numbers refer to the number of repeated thymines and the ampersand represents a special allele class in the homopolymeric region, TTTTTCITTT. All other deletions, insertions, complex mutational events, homopolymeric regions and amino acid changes are described in KIRBY and STEPHAN (1995).

the small introns with $n = 15$ and $\theta = 23.07$ for the whole *white* gene. The probability of observing three or more polymorphisms in intron 2 was $P = 0.043$, of observing three or more polymorphisms in intron 5 was $P = 0.038$, and of observing two or more polymorphisms

in intron 3 was $P = 0.083$. Therefore, it appears that there is a significant excess of silent site variation in introns 2 and 5, given the average nucleotide diversity of the entire *white* gene.

Heterogeneity of fixed differences between *D. melano-*

TABLE 1
Chi-squared tests for heterogeneity of polymorphism and divergence for different domains of the *white* gene

Region	L	Polymorphisms				Divergence		
		S _o	S _e	χ^2	$\hat{\theta}$	F _o	F _e	χ^2
5' UTR	225	6	3.61	1.58	0.0082	3	14.39	9.02
Exon 1	24	0	0.39	0.39	0.0000	0	1.53	1.53
Intron 1	3123	43	50.11	1.01	0.0042	161	184.11	2.90
Exon 2	95	3	1.52	1.44	0.0097	10	6.08	2.67
Intron 2	73	3	1.17	2.86	0.0126	10	4.41	7.09
Exon 3	242	5	3.88	0.32	0.0063	15	15.48	0.01
Intron 3	60	2	0.96	1.13	0.0103	4	3.52	0.07
Exon 4	112	2	1.80	0.02	0.0055	6	7.16	0.19
Intron 4	219	3	3.52	0.08	0.0042	18	12.86	2.05
Exon 5	44	1	0.71	0.12	0.0070	3	2.81	0.01
Intron 5	70	3	1.12	3.16	0.0132	13	4.09	19.41
Exon 6	253	1	4.06	2.31	0.0012	22	16.18	2.09
3' UTR	134	3	2.15	0.34	0.0069	16	8.38	6.93
Total	4674	75	75.00	14.76	0.0049	281	281.00	53.97***
Introns ≤ 90	203	8	3.26	6.89	0.0120	27	12.02	17.88
Introns > 90	3342	46	53.63	1.09	0.0042	179	196.97	2.24
Exons	770	12	12.35	0.01	0.0048	56	49.24	0.71
UTRs	359	9	5.76	1.83	0.0077	19	22.77	0.74
Sum				9.82*				21.87***

S_o is the number of observed silent segregating sites, and S_e is the expected number of segregating sites (which is proportional to the size of the domain). F_o is the observed number of fixed differences between *D. melanogaster* and *D. simulans*, and F_e is the expected number of fixed differences. The domain sizes L (in bp) are based on the reference sequence (Canton-S). θ was estimated according to WATTERSON (1975) and NEI (1987; Equation 10.3). Introns were grouped according to size, large (> 90 bp) and small (≤ 90 bp) (MOUNT *et al.* 1992; STEPHAN *et al.* (1994). * $P < 0.05$; *** $P < 0.001$.

gaster and *D. simulans* within individual domains of the *white* gene was also examined with a goodness of fit test. This test revealed significant heterogeneity of fixed differences between the structural domains of the *white* gene ($\chi^2 = 53.97$; 12 d.f.; $P < 10^{-6}$). An examination of the individual deviations from expected numbers (Table 1) showed that several domains were responsible for the large chi-squared value. Similar to the heterogeneity of silent polymorphism, both intron 2 and intron 5 have a large excess of fixed differences from what is expected given the overall level of divergence for the *white* gene. Furthermore, there appeared to be an excess of fixed differences in the 3' UTR and a deficiency of fixed differences in the 5' UTR. A deficiency of fixed differences within the 5' UTR may not be unusual (LI and GRAUR 1991) and may reflect functional constraints due to transcriptional/translational regulatory functions (BEELMAN and PARKER 1995; CURTIS *et al.* 1995; JOHNSTON 1995). Grouping the domains also resulted in significant heterogeneity in fixed differences, with the highest deviations from expected numbers again coming from the small introns ($\chi^2 = 21.87$, 3 d.f.; $P < 10^{-4}$).

We applied standard neutrality tests over the entire 5972-bp transcriptional unit of the *white* gene. We used the HKA test 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 reference locus.

This resulted in a chi-squared value of 0.30, which is nonsignificant. TAJIMA's (1989) test ($D = -0.07$, NS) and the FU and LI (1993) test ($D^* = 0.29$, NS) both failed to reject the null hypothesis of a neutral equilibrium model as well. The results of these standard neutrality tests applied over the entire transcriptional unit are consistent with previous tests of neutrality applied to a smaller region of the *white* gene (KIRBY and STEPHAN 1995).

Structure of variation within the *white* gene: An examination of Figure 1 indicates that there may be more haplotypic structuring of the silent polymorphisms in the 3' half of the transcriptional unit, as opposed to the 5' half. Differences in the inferred recombinational history between the two halves of the transcriptional unit support this observation. The method of HUDSON and KAPLAN (1985) was used to infer that there were a minimum of 13 recombination events in the history of the sample of 15 *D. melanogaster* sequences. These 13 inferred recombination events occurred in the intervals 3737–3326, 3325–3279, 3209–3121, 3120–3093, 3092–3030, 3029–2968, 2716–2501, 2496–1978, 1723–1509, 1508–1119, 931–245, 244 to –198 and –450 to –2186, where the numbers refer to the nucleotide site positions as in Figure 1. Note that only three of the 13 inferred recombination events occur in the 3352-bp interval from coordinate 1118 to the end of the transcriptional unit. In addition, 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. We calculated the ratio $C:\theta$ for the two halves of the transcriptional unit, and found ratios of 7.73 for the 5' half (coordinates 3738–1119) and 0.10 for the 3' half (coordinates 1118 to –2233), where the average nucleotide diversities per region were estimated to be $\theta = 11.08$ for the 5' half and $\theta = 11.99$ for the 3' half. This may suggest that the haplotypic structuring in the 3' half of *white* is due to low levels of recombination. However, direct measurements of intra-genic recombination in the *white* gene have revealed no evidence for suppression of recombination in the 3' half of the gene (JUDD 1964, 1987; LEFEVER 1973; B. JUDD, personal communication).

We attempted to quantify the structure of variation in the *white* gene by performing significance tests of linkage disequilibrium for all pairs of nonunique silent polymorphisms across the transcriptional unit without correcting for multiple comparisons (SOKAL and ROHLF 1981). Figure 2 shows the results. Although the expected spatial distribution for linkage disequilibria under neutrality is unknown, it is clear from Figure 2 that there are many more nonrandom associations between the polymorphisms in the 3' half of the gene than in the 5' half.

To further quantify the structure of variation within the *white* gene, we used the haplotype test. We applied this test to the entire transcriptional unit of the *white* gene for all possible window sizes (2–61) of nonunique silent polymorphisms without recombination. Figure 3 shows the results for two window sizes. An inspection of the graphs reveals steep drops in the P -values at polymorphism 34, which corresponds to the polymorphism at coordinate 1118. This again indicates that there is much more structuring of the variation in the 3' half of the gene. There appears to be some unevenness to the structuring, however. Although the P -values are much lower in the 3' half of the gene, they actually reach a local maximum at polymorphism 50 in the window size of 5 (polymorphism at coordinate –471) and between polymorphisms 45 and 51 in the window size of 10 (polymorphisms from coordinates 521 to –504), with subsequent decreases near the end of the molecule. These “humps” in the P -values involve the silent polymorphisms that surround the regions where the two amino acid replacement polymorphisms are located. It is therefore possible that the high P -values in this region are caused by hitchhiking of the silent variants in conjunction with the selected dynamics of the replacement polymorphisms.

DNA sequence variation in a 1612-bp segment of the *white* gene in two populations: The 834-bp segment between coordinates 806 and –28 was the only region to have P -values from the haplotype test lower than 5% of the two randomly generated samples. Thus, the haplotype test rejected a neutral equilibrium model as an

explanation for the haplotypic structuring in this segment of the *white* gene. To examine alternative explanations to the neutral equilibrium model, we sequenced a 1612-bp region that included the 834-bp segment for 33 additional lines of *D. melanogaster*: nine from the Beltsville population and 24 from a European population. [Since there is no evidence for differentiation between the populations from Holland and Spain (MARTÍN-CAMPOS *et al.* 1992), these were considered one population.] Polymorphism data for this 1612-bp region are summarized in Figure 4. There was a total of 12 silent segregating sites, five length polymorphisms, two polymorphisms resulting from complex mutations and one replacement polymorphism observed in both samples. In addition, there were five silent segregating sites and one length polymorphism observed only in the European lines, and nine silent segregating sites and four length polymorphisms occurred only in the Beltsville sample.

The polymorphisms within the 834-bp region between coordinates 806 and –28 formed three basic haplotypes: (1) identical to the reference sequence Canton-S (O'HARE *et al.* 1984), frequency: 15/48; (2) differs from Canton-S at one site, which results in a replacement substitution at amino acid residue 49 (PEPLING and MOUNT 1990), frequency: 13/48; and (3) differs from Canton-S by 11 sites, which includes the replacement substitution, frequency: 11/48. To be consistent with KIRBY and STEPHAN (1995), we designated these haplotypes 244T (*e.g.*, XG2), 244G (*e.g.*, XH62) and 244G11 (*e.g.*, XG8), respectively. In addition, nine of the 48 sequences appeared to be recombinant sequences (XH14, XH11, XH19, XH84, XG17, XH44, XH18, Bv30, and Bv15). We used the method of HUDSON and KAPLAN (1985) to infer from our sample the minimum number of recombination events that have occurred in this region. We found two recombination events within the intervals 720–521 and 478–379. Line Bv15 could have resulted from recombination between haplotypes 244T and 244G11, while all other recombinants could have resulted from recombination between haplotypes 244G and 244G11.

An explanation that is consistent with selective neutrality for the high frequency haplotypes within this 834-bp region, as well as the overall haplotypic structuring in the 3' half of the transcriptional unit, is that the haplotypic structuring is due to limited migration between the two subpopulations surveyed. We examined the effect of population structure on the nonrandom associations using OHTA's (1982) method. OHTA (1982) defined variance components of linkage disequilibrium (D_{IS}^2 , D_{ST}^2 , D'_{IS}^2 , D'_{ST}^2 and D'^2_{IT}) that are analogous to WRIGHT's (1940) inbreeding statistics. OHTA (1982) suggested that limited migration is responsible for non-random associations between polymorphisms if $D'^2_{IS} > D'^2_{ST}$ and $D'^2_{ST} > D'^2_{IS}$. In other words, nonrandom associations would not show a consistent pattern between sub-

45 pairwise comparisons, each of which resulted in $D'_{IS} < D'_{ST}$ and $D^2_{ST} < D^2_{IS}$. This suggests that the haplotypes in this region, and the overall haplotypic structuring of variation in the 3' half of the transcriptional unit, cannot be explained by a neutral model with limited migration between the European and North American populations.

KIRBY and STEPHAN (1995) proposed two selective explanations for the high frequency haplotypes in the 834-bp region. The first hypothesis entails directional

FIGURE 3.—*P*-values for sliding windows of the haplotype test across the transcriptional unit of the *white* gene. The nonunique silent polymorphisms are numbered along the *x*-axis. Polymorphism number 34 corresponds to the polymorphism at coordinate 1118 and polymorphism number 39 corresponds to the polymorphism at coordinate 806. (a) Window size of five polymorphisms. (b) Window size of 10 polymorphisms.

and/or recent balancing selection. Since the predictions for these mechanisms are similar (see HUDSON *et al.* 1994), we refer to these forces in the following jointly as directional (or positive) selection. The second hypothesis involves epistatic fitness interactions between the *white* polymorphisms. Two criteria were proposed to distinguish between these two selection hypotheses: (1) the amount of recombination between the haplotypes, and (2) the amount of nucleotide diversity within each haplotype relative to the average nucleotide diversity within the region. To examine these criteria, it is appropriate to group the haplotypes 244G and 244T into a single haplotype, 244G/244T (KIRBY and STEPHAN 1995). In contrast to the directional selection model, the epistatic selection hypothesis predicts

that recombinants between 244G/244T and 244G11 are likely to be eliminated and thus would be in low frequency. With regard to the second criterion the epistatic selection hypothesis predicts low variation within all haplotype classes, while the directional selection hypothesis predicts low variation only in the haplotype class that is of relatively recent origin.

We examined the amount of recombination between the haplotype classes in the 834-bp segment using an estimator of the recombination parameter $C = 3Nc$ (HUDSON 1987). This estimator, $\hat{C}$, is based on the variance in the number of pairwise differences between the sequences in a sample. Recombinants present in the sample increase this variance, resulting in larger values of $\hat{C}$. An estimate for C using HUDSON's (1987) method

a.		b.	
I		I	
1		1	
a		i	
b c		j k l	
d e f		m n	
g h			
1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 -		1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -	
5 4 3 1 0 0 9 9 8 8 7 7 7 6 5 4 3 2 0		3 1 1 1 9 7 6 5 5 4 4 3 1 1 1 1 0 0 0 0	
0 0 9 1 6 4 5 3 0 0 7 4 2 4 2 7 7 4 2		1 5 5 4 6 3 2 5 5 4 0 2 5 4 3 1 9 1 1 0 2	
8 7 7 8 6 5 5 1 6 1 9 4 0 9 1 8 9 4 8		0 4 4 4 8 6 2 6 5 5 4 3 6 4 5 3 6 8 4 1 2	
CanS A * G A * * A C T C T * * * A G * T T		CanS T G A * T * A * * C T A T T C A A A G * *	
XG8 T M C . . . C T C G C D D M G T I G C		XG8 . D	
XH61 T M C . I D C T C G C D D M G T I G C		XH61 . T .	
XH30 T M C . I D C T C G C D D M G T I G C		XH30 . T . D	
XH1 T M C . I D C T C G C D D M G T I G C		XH1 . T .	
XH83 T M C . I D C T C G C D D M G T I G C		XH23 .	
XH23 .		XH36 .	
XH14 .		XH11 .	
XH11 .		XG17 .	
XH19 .		XH118 .	
XH84 .		BV13 .	
XG17 .		BV10 .	
XH44 .		BV7 .	
XH62 T M C		BV16 .	
XG30 .		BV23 .	
XH18 .		BV5 C .	
XG67 .		BV3 C .	
XG2 T M C		BV6 .	
XH48 .		BV30 .	
XG10 .		BV14 .	
XH118 .		BV39 .	
XH55 .		BV19 .	
XH34 .			
XH36 .			
XH3 .			
BV10 T .			
BV7 T M C			
BV16 T M C			
BV23 . M C T I D C T C G C D D M G T I G C			
BV5 T M C . I D C T C G C D D M G T I G C			
BV30 .			
BV13 .			
BV2 .			
BV4 .			
BV8 T M C			
BV31 T M C			
BV22 .			
BV37 .			
BV9 .			
BV32 .			
BV6 .			
BV3 .			
BV38 .			
BV15 .			
BV19 .			
BV12 .			
BV14 .			
BV39 .			
BV36 .			

FIGURE 4.—DNA polymorphisms in a 1612-bp segment of *white* in two populations of *D. melanogaster*. Lines with the prefix XH are from Spain, lines with the prefix XG are from Holland and lines with the prefix Bv are from Maryland. (a) Polymorphisms with observed frequencies of the rare allele $\geq 10\%$. (b) Polymorphisms found with observed frequencies of the rare allele $< 10\%$. Symbols and numbering are the same as in Figure 1. Insertions, deletions, complex mutations and amino acid changes are as follows: a, 1407 to 1402 replaced by TTTTTTTTTTTT; b, TAG; c, 1045 to 1032; d, 744 to 739; e, 720 to 714; f, 649 to 643 replaced by CTTTTT; g, GGG-TACT; h, Arg (CGG) to Leu (CUG); i, 1144; j, 736 to 730; k, 556; l, G; m, -1 to -2; n, -22.

for the region between positions 806 and -28 is $\hat{C} = 1.42$, excluding length polymorphisms. This estimate is for the 18 silent segregating sites in this region, which includes the nucleotide site polymorphism resulting from a complex mutational event. To better understand this estimate, we generated a null distribution for $\hat{C}$ by randomly permuting the columns of the observed data. This created a new set of sequences with the same set of base frequencies as the observed data, but destroyed the spatial ordering. We calculated $\hat{C}$ for 1000 permuted samples, none of which showed a lower value than the observed one. In fact, none of the permuted samples resulted in $\hat{C}$ values lower than 8.00. As a comparison the region between coordinates 1508 and 931, which comprises nine segregating sites, gave an estimate of $\hat{C} = 8.20$. Although a low $\hat{C}$ value provides evidence for epistatic selection, it also raises the possibility that the haplotype structures are due to a local reduc-

tion in recombination rate. However, the estimator $\hat{C}$ is based on the variance in the numbers of pairwise differences between the sequences in a sample, which means that $\hat{C}$ estimates the frequency of recombinants in the sample rather than the actual recombination rate. Therefore, it is possible to have a high recombination rate, while observing low frequencies of recombinants. We also inferred the minimum number of recombination events that have occurred within the 834-bp region using the method of HUDSON and KAPLAN (1985). This method is based upon observing the four possible genotypic classes in a sample (for a two-locus, two-allele case). We inferred a minimum of two recombination events in the sample of 48 sequences within the intervals 720–521 and 478–379. This means that although there are few recombinants present in the sample, there is evidence of recombination. Furthermore, direct measurement of intragenic recombination

rates suggests that there is no suppression of recombination in this segment of the *white* gene (JUDD 1964, 1987; LEFEVER 1973). All these observations support an epistatic selection hypothesis for this region, and possibly for the overall structuring in the 3' half of the gene.

With regard to the second criterion, we estimated the amount of nucleotide diversity within each haplotype relative to the average nucleotide diversity within the region between coordinates 806 and -28. We found that 18 silent sites are segregating in this region (out of a total of 659 silent sites). The estimate of the average nucleotide diversity, $\hat{\theta} = 0.0061$, is slightly larger than the estimate for the entire transcriptional unit (Table 1). We used TAVARÉ's (1984) method to calculate for each haplotype class the probability, P , of having j segregating sites in a sample of size n , given that average nucleotide diversity per region is $\theta = 659 \times 0.0061 = 4.02$. For haplotype class 244T/244G (with $n = 28$, $j = 3$), the probability of observing three or fewer segregating sites is $P = 0.0041$; for haplotype class 244G11 ($n = 11$, $j = 8$), the probability of observing eight or fewer segregating sites is $P = 0.33$. The low nucleotide diversity in haplotype class 244T/244G supports both the epistatic selection and the directional selection hypotheses. However, nucleotide diversity in haplotype class 244G11 is not significantly lower than average nucleotide diversity for the entire 659-bp region, which provides support for the directional selection hypothesis.

DISCUSSION

Overview: We surveyed nucleotide variation for the entire transcriptional unit of the *white* gene of a *D. melanogaster* population from Maryland. Differences in the inferred recombinational histories suggested haplotypic structuring of silent variation in the 3' half of the *white* gene, as opposed to the 5' half. The distribution of linkage disequilibria and the haplotype test confirmed this result. In fact, there are three basic haplotypes among the 15 *D. melanogaster* lines for the silent polymorphisms from coordinate 1118 to the end of the transcriptional unit, whereas the polymorphisms from the beginning of the transcriptional unit to coordinate 1119 form 15 distinct haplotypes. For the following discussion of these results, we designate the haplotypes in the entire 3' half as haplotype A (similar to Bv 2), haplotype B (similar to Bv 6) and haplotype C (Bv 7 and Bv 10); in addition, we have Bv 8, which appears to be a recombinant between haplotypes B and C.

Such a drastic difference between the two halves of the transcriptional unit may suggest that the haplotypic structuring is due to differences in recombination rate; however, there is no evidence for suppression of recombination in the 3' half of the gene (JUDD 1964, 1987; LEFEVER 1973). Furthermore, we found that the haplotypic structuring is not consistent with a neutral equilibrium model of random mutation and genetic drift. To

examine whether natural selection is responsible for the haplotypic structuring, we sequenced a 1612-bp region of the *white* gene for an additional 33 lines of *D. melanogaster*, 24 of which are from a European population. Using these data, we specifically investigated the two selection hypotheses that were proposed by KIRBY and STEPHAN (1995). The first hypothesis entails directional selection or recent balancing selection within or near the *white* gene and the second hypothesis involves epistatic fitness interactions between the polymorphic sites. We examined two criteria to distinguish between these two selection hypotheses: the frequency of recombinants within the region, and the amount of nucleotide diversity within each haplotype relative to the average nucleotide diversity within the region. Analyses of these two criteria for the 48 *D. melanogaster* sequences did not provide clear evidence against either of the two selection hypotheses. An examination of these same two criteria for the entire 3' half of the gene showed similar results. There was a low frequency of recombinants within the 3' half of the gene, which supports an epistatic selection hypothesis for the haplotypic structuring. The estimated C: θ ratio of 0.10 for the 3' half was exceptionally low for a gene that is not located in a region of low recombination. As a comparison, in the *Adh* region the C: θ ratio was estimated by HUDSON (1987) to be 1.6. On the other hand, an analysis of nucleotide diversity within each of the haplotypes in the 3' half relative to the average nucleotide diversity for the whole 3' half provided support for a directional selection hypothesis for the haplotypic structuring. In fact, when TAVARÉ's (1984) method is applied to the three haplotypes in the 3' half of the gene (with $\theta = 11.99$ for the entire 3' half; see above), we find the following: For haplotype A ($n = 4$, $j = 0$), the probability of observing zero segregating sites is $P = 0.002$; for haplotype B ($n = 8$, $j = 8$), the probability of observing eight or fewer segregating sites is $P = 0.02$; for haplotype C ($n = 2$, $j = 14$), the probability of observing 14 or fewer segregating sites is $P = 0.711$. (The number of segregating sites within haplotypes B and C include one and two nucleotide differences within homopolymeric tracts, respectively.) This suggests that haplotype A (and perhaps also B) have recently risen into high frequency due to positive selection.

An evolutionary model: Our analyses suggest that the *white* polymorphism data are incompatible with either a pure epistatic selection or a pure directional/balancing selection explanation for the haplotypic structuring in the 3' half. Therefore, we propose a model that combines features of these two hypotheses and takes into account the recent history of *D. melanogaster*. Several studies have suggested that populations of *D. melanogaster* have recently migrated from older populations in Africa to the rest of the world (BEGUN and AQUADRO 1993, 1995). We propose that during this process positive Darwinian selection acted at several different sites

in the 3' half or downstream of the *white* gene in response to local adaptation to new environments. Haplotypes entering new environments were thus drawn quickly into intermediate or high frequency; however, they were prevented from going to fixation due to "traffic" with haplotypes that were targeted by selection at different sites. A quantitative analysis that describes this traffic model for nucleotide variation is currently not available. STEPHAN (1995) studied a two-locus, two-allele model with directional selection at both loci, assuming that the effects of selection were additive. For the case that the selection coefficients for both loci are comparable and recombination between the loci is infrequent (which appears to be appropriate for intragenic recombination), his analysis suggests that it may take a long time until the favorable alleles at both loci are combined into a single haplotype (with higher fitness) that is then eventually being fixed. This means that neutral or nearly neutral variation can accumulate on the selected haplotypes before more favorable combinations arise and go to fixation.

Do the *white* polymorphism data support such a model? One prediction of this model is that there may be differences in the ages of the haplotypes. This appears to be the case for the haplotypes in the 3' half of the *white* gene. Comparison with the *D. simulans* sequence and a high nucleotide diversity indicate that haplotype C is much older than haplotypes A and B. There are 32 nucleotide sites in the 3' half of the *white* gene that are variable within the *D. melanogaster* sample from Maryland and homologous between *D. melanogaster* and *D. simulans*. Of these 32 nucleotide sites, 28 are identical to one of the two sequences comprising haplotype C. In addition, the nucleotide diversity per site of haplotype C based on the average number of pairwise differences is estimated to be $\hat{\pi} = 0.0067$, with 14 segregating sites between the two sequences. Such a large accumulation of nucleotide polymorphisms indicates that haplotype C has been around for a long time. In contrast, haplotypes A and B have relatively low levels of nucleotide diversity, suggesting that these haplotypes have arrived recently in the population we sampled. The nucleotide diversity per site of haplotype B based on the average number of pairwise differences is $\hat{\pi} = 0.0011$, with eight segregating sites between the eight sequences, and the four sequences of haplotype A contain no segregating sites at all. The apparent age differences between the three haplotypes support a recent migration of haplotypes into the population we sampled.

However, there may be evidence against a simple traffic model with additive selection. We noticed a relatively high level of nucleotide diversity in one of the "new" haplotypes, haplotype B, suggesting that this haplotype has been around long enough to accumulate molecular variation. This raises the question of why this haplotype has not been broken up by recombination, given that

it is relatively old and that levels of recombination in the 3' half of *white* are not reduced. In fact, the number of recombination events observed in the entire 3' half of *white* seems to be very low. This may imply that a more realistic traffic model should involve epistatic interactions between polymorphisms as well. A quantitative analysis of the proposed model is needed to examine this possibility.

Our multi-locus selection model may also explain the geographic pattern of the haplotypic structuring that has been observed at other loci in *D. melanogaster*. For instance, extensive haplotype structures have been found at the *vermillion* locus (BEGUN and AQUADRO 1995) in non-African populations, but African populations from the putative zoogeographic center of this species exhibit only small levels of linkage disequilibrium. The haplotypic structures in the *Sod* region, for which only data from non-African populations are available (HUDSON *et al.* 1994), may be interpreted in a similar way. Surveys of African populations for both *white* and *Sod* may shed new light on the proposed model.

We thank MONTERRAT AGUADÉ, CHUCK LANGLEY, DAVE BEGUN and two anonymous reviewers for their comments and suggestions on this manuscript. We also thank MONTERRAT AGUADÉ for sharing European fly lines. We would like to acknowledge JEB GAITHER and JOHN PARSCH for help with computer programming. This research was supported by a grant from the National Science Foundation (DEB-9407226) and a Semester Research Award from the University of Maryland to W.S.

LITERATURE CITED

- ASMUSSEN, M. A., and M. T. CLEGG, 1981 Dynamics of the linkage disequilibrium function under models of gene-frequency hitchhiking. *Genetics* **99**: 337–356.
- BEELMAN, C. A., and R. PARKER, 1995 Degradation of mRNA in eukaryotes. *Cell* **81**: 179–183.
- 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 *vermillion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. *Genetics* **140**: 1019–1032.
- 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.
- CURTIS, D., R. LEHMANN and P. D. ZAMORE, 1995 Translational regulation in development. *Cell* **81**: 171–178.
- 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.
- HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**: 226–231.
- HUDSON, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**: 611–631.
- 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.
- JOHNSTON, D. S., 1995 The intracellular localization of messenger RNAs. *Cell* **81**: 161–170.
- JUDD, B. H., 1964 The structure of intralocus duplication and deficiency chromosomes produced by recombination in *Drosophila melanogaster*, with evidence for polarized pairing. *Genetics* **49**: 253–265.
- 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. HENNIG. Springer-Verlag, Berlin.
- KIMURA, M., 1956 A model of a genetic system which leads to closer linkage by natural selection. *Evolution* **10**: 278–287.
- KIRBY, D. A., and W. STEPHAN, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. *Genetics* **141**: 1483–1490.
- 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.
- 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.
- LAURIE, C. C., and L. F. STAM, 1994 The effect of an intronic polymorphism on alcohol dehydrogenase expression in *Drosophila melanogaster*. *Genetics* **138**: 379–385.
- LAURIE, C. C., J. T. BRIDGHAM and M. CHOUDHARY, 1991 Associations between DNA sequence variation and variation in expression of the *Adh* gene in natural populations of *Drosophila melanogaster*. *Genetics* **129**: 489–499.
- LAURIE-AHLBERG, C. C., 1985 Genetic variation affecting the expression of enzyme-coding genes in *Drosophila*: an evolutionary perspective. *Isozymes Curr. Top. Biol. Med. Res.* **12**: 33–88.
- LEFEVER, H. M., 1973 Analysis of three white mutants resulting in two new recombination sites at the *white* locus in *Drosophila melanogaster*. *Dros. Inf. Ser.* **50**: 109–110.
- LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.
- LEWONTIN, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. *Genetics* **140**: 377–388.
- LI, W.-H., and D. GRAUR, 1991 *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, MA.
- MARTÍN-CAMPOS, J. M., J. M. COMERÓN, N. MIYASHITA and M. AGUADÉ, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics* **130**: 805–816.
- MERRIAM, J. R., 1968 New mutants report. *Dros. Inf. Ser.* **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. Camb.* **62**: 101–109.
- MOUNT, S. M., C. BURKS, G. D. STORMO, J. HERTZ and C. FIELDS, 1992 Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* **20**: 4255–4262.
- NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- ODGERS, W. A., M. J. HEALY and J. G. OAKESHOTT, 1995 Nucleotide polymorphism in the 5' promoter region of esterase 6 in *Drosophila melanogaster* and its relationship to enzyme activity variation. *Genetics* **141**: 215–222.
- 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.
- OHTA, T., and M. KIMURA, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**: 229–238.
- 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.
- SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry*. W. H. Freeman and Co., New York.
- STEPHAN, W., 1995 Perturbation analysis of a two-locus model with directional selection and recombination. *J. Math. Biol.* **34**: 95–109.
- STEPHAN, W., and S. J. MITCHELL, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. *Genetics* **132**: 1039–1045.
- STEPHAN, W., V. S. RODRIGUEZ, B. ZHOU and J. PARSCH, 1994 Molecular evolution of the metallothionein gene *Mtn* in the *melanogaster* species group: results from *Drosophila ananassae*. *Genetics* **138**: 135–143.
- STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**: 149–153.
- TACHIDA, H., 1994 Decay of linkage disequilibrium in a finite island model. *Genet. Res. Camb.* **64**: 137–144.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- TALERICO, M., and S. M. BERGET, 1994 Intron definition in splicing of small *Drosophila* introns. *Mol. Cell Biol.* **14**: 3434–3445.
- TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. *Theor. Popul. Biol.* **26**: 119–164.
- WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**: 256–276.
- WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. *Am. Nat.* **74**: 232–248.

Communicating editor: D. CHARLESWORTH