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 (Odgars 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 populations of D. melanogaster at the vermilion 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-Ahlgren 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.
MATERIALS AND METHODS

Drosophila lines and DNA sequencing: Nine isofemale \textit{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 \textit{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, XH184 and XH118) and Groningen, Holland (XG2, XG8, XG10, XG17, XG30 and XG67), were previously described in MARTIN-CAMPOS et al. (1992). Fifteen lines of \textit{D. melanogaster} (Bv2, Bv3, Bv4, Bv6, Bv7, Bv8, Bv10, Bv12, Bv14, Bv19, Bv22, Bv36, Bv37, Bv38, Bv39) and one \textit{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. 1984).

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 \textit{D. melanogaster} lines and for the \textit{D. simulans} line used in KIRBY and STEPHAN (1995). The primers for each segment, corresponding to coordinates from the previously published \textit{D. melanogaster} white sequence (O'HAIRE et al. 1984), were as follows: (internal primers are in parentheses): 3791 -3772 (3764-XG67), were previously described in MARTIN-CAMPOS et al. (1992). The sequences were tested for departures from selective neutrality using the HUDSON, KREITMAN and AGUADE (HKA) 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 \( \pi \) and \( \theta \) for the 15 \textit{D. melanogaster} sequences are as follows: \( \pi = 0.0049; \theta = 0.0049 \). 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 (\( \chi^2 = 14.76; 12 \text{ d.f.; NS} \)). Although this test failed to reveal evidence for between-domain heterogeneity, there were several domains whose estimated \( \theta \)'s were much higher than the estimated \( \theta \) 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 (\( \chi^2 = 9.82; 3 \text{ d.f.; } P < 0.025 \)). 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 TAVARE (1984). For a zero-recombination model, TAVARE (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, \( \theta \) (per region). We calculated \( P \) for each of
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 \( D. melanogaster \)
**TABLE 1**

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

<table>
<thead>
<tr>
<th>Region</th>
<th>L</th>
<th>S₀</th>
<th>S₁</th>
<th>χ²</th>
<th>θ</th>
<th>F₀</th>
<th>F₁</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' UTR</td>
<td>225</td>
<td>6</td>
<td>3.61</td>
<td>1.58</td>
<td>0.0082</td>
<td>3</td>
<td>14.39</td>
<td>9.02</td>
</tr>
<tr>
<td>Exon 1</td>
<td>24</td>
<td>0</td>
<td>0.39</td>
<td>0.39</td>
<td>0.0000</td>
<td>0</td>
<td>1.53</td>
<td>1.53</td>
</tr>
<tr>
<td>Intron 1</td>
<td>3123</td>
<td>43</td>
<td>50.11</td>
<td>1.01</td>
<td>0.0042</td>
<td>161</td>
<td>184.11</td>
<td>2.90</td>
</tr>
<tr>
<td>Exon 2</td>
<td>95</td>
<td>3</td>
<td>1.52</td>
<td>1.44</td>
<td>0.0007</td>
<td>10</td>
<td>6.08</td>
<td>2.67</td>
</tr>
<tr>
<td>Intron 2</td>
<td>73</td>
<td>3</td>
<td>1.17</td>
<td>2.86</td>
<td>0.0126</td>
<td>10</td>
<td>4.41</td>
<td>7.09</td>
</tr>
<tr>
<td>Exon 3</td>
<td>242</td>
<td>5</td>
<td>3.88</td>
<td>0.32</td>
<td>0.0063</td>
<td>15</td>
<td>15.48</td>
<td>0.01</td>
</tr>
<tr>
<td>Intron 3</td>
<td>60</td>
<td>2</td>
<td>0.96</td>
<td>1.13</td>
<td>0.0103</td>
<td>4</td>
<td>9.52</td>
<td>0.07</td>
</tr>
<tr>
<td>Exon 4</td>
<td>112</td>
<td>2</td>
<td>1.80</td>
<td>0.02</td>
<td>0.0055</td>
<td>6</td>
<td>7.16</td>
<td>0.19</td>
</tr>
<tr>
<td>Intron 4</td>
<td>219</td>
<td>3</td>
<td>3.52</td>
<td>0.08</td>
<td>0.0042</td>
<td>18</td>
<td>12.86</td>
<td>2.05</td>
</tr>
<tr>
<td>Exon 5</td>
<td>44</td>
<td>1</td>
<td>0.71</td>
<td>0.12</td>
<td>0.0070</td>
<td>3</td>
<td>2.81</td>
<td>0.01</td>
</tr>
<tr>
<td>Intron 5</td>
<td>70</td>
<td>3</td>
<td>1.12</td>
<td>3.16</td>
<td>0.0132</td>
<td>13</td>
<td>4.09</td>
<td>19.41</td>
</tr>
<tr>
<td>Exon 6</td>
<td>253</td>
<td>1</td>
<td>4.06</td>
<td>2.31</td>
<td>0.0012</td>
<td>22</td>
<td>16.18</td>
<td>2.09</td>
</tr>
<tr>
<td>3' UTR</td>
<td>134</td>
<td>3</td>
<td>2.15</td>
<td>0.34</td>
<td>0.0069</td>
<td>16</td>
<td>8.38</td>
<td>6.93</td>
</tr>
<tr>
<td>Total</td>
<td>4674</td>
<td>75</td>
<td>75.00</td>
<td>14.76</td>
<td>0.0049</td>
<td>281</td>
<td>281.00</td>
<td>53.97***</td>
</tr>
<tr>
<td>Introns ≤90</td>
<td>203</td>
<td>8</td>
<td>3.26</td>
<td>6.89</td>
<td>0.0120</td>
<td>27</td>
<td>12.62</td>
<td>17.88</td>
</tr>
<tr>
<td>Intron &gt;90</td>
<td>3342</td>
<td>46</td>
<td>53.63</td>
<td>1.09</td>
<td>0.0042</td>
<td>179</td>
<td>196.97</td>
<td>2.24</td>
</tr>
<tr>
<td>Exons</td>
<td>770</td>
<td>12</td>
<td>12.55</td>
<td>0.01</td>
<td>0.0048</td>
<td>56</td>
<td>49.24</td>
<td>0.71</td>
</tr>
<tr>
<td>UTRs</td>
<td>359</td>
<td>9</td>
<td>5.76</td>
<td>1.83</td>
<td>0.0077</td>
<td>19</td>
<td>22.77</td>
<td>0.74</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td>9.82*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21.87***</td>
</tr>
</tbody>
</table>

S₀ is the number of observed silent segregating sites, and S₁ is the expected number of segregating sites (which is proportional to the size of the domain). F₀ is the observed number of fixed differences between D. melanogaster and D. simulans, and F₁ 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.

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 (Kirkby 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 \( C:0 \) 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 esti-
imated 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; Leefever 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 50) 
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 mole-
cule. These “humps” in the \( P \)-values involve the silent 
polymorphisms that surround the regions where the 
two amino acid replacement polymorphisms are lo-
eated. It is therefore possible that the high \( P \)-values in 
this region are caused by hitchhiking of the silent vari-
ants 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 haplo-
type test rejected a neutral equilibrium model as an 
exploration for the haplotypic structuring in this segment 
of the \( white \) gene. To examine alternative explana-
tions 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 popula-
tion. [Since there is no evidence for differentiation be-
tween the populations from Holland and Spain (Mar-
tin-Campos et al. 1992), these were considered one popu-
lation.] Polymorphism data for this 1612-bp region 
are summarized in Figure 4. There was a total of 
defined three high frequency 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 replace-
ment 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 replace-
ment 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 se-
quences (XH14, XH11, XH19, XH84, XG17, XH44, 
XH18, Bv30, and Bv15). We used the method of Hud-
don and Kaplan (1985) to infer from our sample the 
minimum number of recombination events that have 
ocurred 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 recombi-
ants could have resulted from recombination between 
haplotypes 244G and 244G11.

An explanation that is consistent with selective neu-
trality 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 be-
tween the two subpopulations surveyed. We examined 
the effect of population structure on the nonrandom 
declared variance components of linkage disequilibrium 
(\( D_{ST}^2 \), \( D_{ST}^2 \), \( D_{ST}^2 \), \( D_{ST}^2 \) and \( D_{ST}^2 \)) that are analogous to 
Wright's (1940) inbreeding statistics. Ohta (1982) 
suggested that limited migration is responsible for non-
random associations between polymorphisms if \( D_{ST}^2 > D_{ST}^2 \) 
and \( D_{ST}^2 > D_{ST}^2 \). In other words, nonrandom associa-
tions would not show a consistent pattern between sub-

populations with regard to strength and direction of association if limited migration were responsible. It has been suggested that low-frequency alleles should not be included in analyses of linkage disequilibria (LEWONTIN 1995). Therefore, we only included polymorphisms with observed frequencies of the rare allele ≥10% in this analysis (HUDSON 1985). We applied Ohta’s (1982) method to the 10 silent polymorphisms with observed frequencies of the rare allele ≥10% for the region between coordinates 806 and −28. There were a total of 45 pairwise comparisons, each of which resulted in $D'_{ST} < D''_{ST}$ and $D''_{ST} < D_{ST}$. 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
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
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 comprise 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 reduction 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 of 48 sequences within the sample, there is evidence of recombination. Furthermore, direct measurement of intragenic recombination

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 ≥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 TTTTTTTTTTT; b, TAG; c, 1045 to 1032; d, 744 to 739; e, 720 to 714; f, 649 to 643 replaced by CTTTTT; g, GGGTACT; h, Arg (CGG) to Leu (CUG); i, 1144; j, 736 to 720; k, 556; l, G; m, -1 to -2; n, -22.
rates suggests that there is no suppression of recombination in this segment of the white gene (JUDD 1964, 1987; LEFEEVER 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 TAVARE’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; LEFEEVER 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 \( C/\theta \) 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 \( C/\theta \) 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 5' half provided support for a directional selection hypothesis for the haplotypic structuring. In fact, when TAVARE’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 vermilion 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 MONTSERRAT AGUADE, CHUCK LANGLEY, DAVE BEGUN and two anonymous reviewers for their comments and suggestions on this manuscript. We also thank MONTSERRAT AGUADE for sharing European fly lines. We would like to acknowledge JEB GAITHER and JOHN PARSH 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


