Evidence for Positive Selection in the Superoxide Dismutase (*Sod*) Region of *Drosophila melanogaster*

Richard R. Hudson, Kevin Bailey, Douglas Skarecky, Jan Kwiatowski and Francisco J. Ayala

Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717

Manuscript received June 21, 1993
Accepted for publication December 6, 1993

ABSTRACT

DNA sequence variation in a 1410-bp region including the Cu,Zn *Sod* locus was examined in 41 homozygous lines of *Drosophila melanogaster*. Fourteen lines were from Barcelona, Spain, 25 were from California populations and the other two were from laboratory stocks. Two common electromorphs, SOD^S and SOD^F, are segregating in the populations. Our sample of 41 lines included 19 Sod^S and 22 Sod^F alleles (henceforward referred to as Slow and Fast alleles). All 19 Slow alleles were identical in sequence. Of the 22 Fast alleles sequenced, nine were identical in sequence and are referred to as the Fast A haplotypes. The Slow allele sequence differed from the Fast A haplotype at a single nucleotide site, the site that accounts for the amino acid difference between SOD^S and SOD^F. There were nine other haplotypes among the remaining 13 Fast alleles sequenced. The overall level of nucleotide diversity (π) in this sample is not greatly different than that found at other loci in *D. melanogaster*. It is concluded that the Slow/Fast polymorphism is a recently arisen polymorphism, not an old balanced polymorphism. The large group of nearly identical haplotypes suggests that a recent mutation, at the *Sod* locus or tightly linked to it, has increased rapidly in frequency to around 50%, both in California and Spain. The application of a new statistical test demonstrates that the occurrence of such large numbers of haplotypes with so little variation among them is very unlikely under the usual equilibrium neutral model. We suggest that the high frequency of some haplotypes is due to natural selection at the *Sod* locus or at a tightly linked locus.

IN recent years, the Cu,Zn superoxide dismutase (SOD) enzyme has been the focus of a number of studies in several organisms, including humans and Drosophila. SOD has been implicated in human disease (*e.g.*, CEBALLOS *et al.* 1990; ROSEN *et al.* 1993), and its possible role in aging has been much discussed (*e.g.*, CEBALLOS-PICOT *et al.* 1992; TANIGUCHI 1992). In *Drosophila melanogaster*, the *Sod* locus is on chromosome *3*, and codes for a homodimeric metalloenzyme, each unit of which is 151 amino acids long in the functional enzyme (LEE *et al.* 1985) and clearly homologous to the human SOD enzyme. It is known from electrophoretic surveys that many natural populations of *D. melanogaster* are segregating for two common alleles (Sod^S and Sod^F) at the *Sod* locus (SINGH *et al.* 1982; PENG *et al.* 1986). (Henceforward, the Sod^S and Sod^F alleles are referred to as the Slow and Fast alleles, respectively.) SOD^S differs from SOD^F by a single amino acid residue, having a lysine instead of an asparagine at residue 96 (LEE and AYALA 1985). LEE *et al.* (1981) purified SOD^S and SOD^F and found that they differ in properties such as specific activity and thermostability (see also GRAF and AYALA 1986). Laboratory experiments have shown that the Slow and Fast alleles, or variation in linkage disequilibrium with them, have large fitness effects in the presence of ionizing radiation (PENG *et al.* 1986), as well as under different conditions of temperature and larval crowding (PENG *et al.* 1991). Recent experiments with laboratory populations of *D. melanogaster* suggest that variation at the *Sod* locus, or a closely linked locus, may be involved in aging in *D. melanogaster* (unpublished data from MICHAEL ROSE's and our laboratory). All of these experiments imply that variation at *Sod*, or at tightly linked sites, has important phenotypic effects on which natural selection may act. Such variation is unlikely to persist for very long unless some form of diversity preserving selection is operating. Examples of such diversity preserving selection are frequency dependent selection, heterozygote superiority, certain forms of temporal and spatial fluctuations of selection coefficients, as well as spatial variation in selection coefficients with limited gene flow.

To clarify the role of natural selection in shaping the polymorphism at the *Sod* locus, we have examined the nucleotide variation at the *Sod* locus in natural populations of *D. melanogaster*. This approach was motivated by a number of recent studies that have used patterns of polymorphism and/or divergence at the DNA level to make inferences about the recent evolutionary histories of particular genes or molecules (*e.g.*, HUDSON and KAPLAN 1988; HUGHES and NEI 1988; KREITMAN and HUDSON 1991; MCDONALD and KREITMAN 1991; EANES *et al.* 1993). These studies have provided support for the role of natural selection, either in maintaining polymorphism, or producing fixations of advantageous mutations. Population genetics theory has shown that a nucleotide polymorphism that has been maintained for a long time by balancing selection, can result in the accumu-

lation of variation at tightly linked sites (HUDSON and KAPLAN 1988). Thus a signature of an old balanced polymorphism can appear in the pattern of sequence variation within a species. The accumulation of variation occurs only at sites very tightly linked to the site at which the balancing selection occurs. Thus the site at which balancing selection has been acting, can in principle be localized to a very small region by an examination of variation in samples from natural populations.

We examined the DNA sequence variation in a 1410-bp region including the coding region of *Sod*, in 41 lines of *D. melanogaster*. The lines came from localities in California and in Barcelona, Spain, and included 19 Slow alleles and 22 Fast alleles. This sampling of approximately equal numbers of Slow and Fast alleles was done so that the level of variation within alleles could be effectively compared to the level of divergence between alleles. If the Slow/Fast polymorphism is an old balanced polymorphism, simple models predict a large divergence between alleles in the neighborhood of the site at which selection acts.

We found that the Slow/Fast polymorphism is not an old polymorphism. In fact, all 19 sequences of Slow alleles were found to be identical in DNA sequence for the entire 1410-bp region examined, and to differ from the most common Fast haplotype by a single nucleotide, the nucleotide that accounts for the amino acid difference between the Fast and Slow forms of the enzyme. Although the Slow/Fast polymorphism is apparently not an old balanced polymorphism, there is evidence in the pattern of variation for the recent action of natural selection. A group of haplotypes, which together have a frequency near one-half in populations of both California and Spain, have very little nucleotide variation within the group, while the rest of the haplotypes have typical *D. melanogaster* levels of variation. Using a new statistical test of neutrality, it is demonstrated that there is too little variation within a subset of the sample, given the level of variation in the rest of the sample. The pattern of variation suggests that a variant at the *Sod* locus or a tightly linked locus has recently risen rapidly in frequency, due to natural selection.

MATERIALS AND METHODS

Sampling from localities: To estimate the frequency of the Slow and Fast alleles, large samples of flies were obtained from two California populations separated by 650 km, Culver City (Los Angeles County) and El Rio vineyard (Lockeford, San Joaquin County) and from Barcelona (Freixenet Winery), Spain. The localities, dates and sample sizes are shown in Table 1. The *Sod* genotypes were ascertained by starch gel electrophoresis and selective staining (AYALA *et al.* 1972). Flies homozygous by descent for the complete third chromosome were obtained by crosses with balancer stocks (SEAGER and AYALA 1982).

Small samples of Slow and Fast alleles were amplified, cloned and sequenced by the methods described below. The samples for DNA sequencing were non-random with respect to the Slow and Fast alleles. To have reasonable numbers of Slow alleles, to estimate within-allele variation, as well as between Slow and Fast variation, approximately equal numbers of Slow and Fast alleles were sequenced from each locality. From Culver City 10 Slow alleles and 9 Fast alleles were sampled for sequencing. From Barcelona, 5 Slow and 9 Fast alleles were sampled. From El Rio, 3 Slow alleles and 2 Fast alleles were sequenced. A Slow allele from Davis (Sacramento County), California was also sequenced. The sequence of *Sod* from the Canton-S strain was also obtained. This is a Fast allele. The sequence of *Sod* of an Oregon-R strain (SETO *et al.* 1989) is also included in some parts of the analysis.

DNA preparation, amplification, cloning and sequencing: Genomic DNA was isolated from 10 to 20 flies by the method of KAWASAKI (1990). The *Sod* gene region was amplified by the polymerase chain reaction (PCR) technique (SAIKI *et al.* 1988) using high fidelity conditions (KWIATOWSKI *et al.* 1991). In short, the reaction mix containing 1–5 μg DNA, PCR buffer (Perkin-Elmer Cetus), 0.2 μM of each primer, 1.5 mM $MgCl_2$ and 40 μM of each dNTP was heated at 95° for 10 min then cooled to 72°. AmpliTaq (2.5 units) was added to each sample along with 100 μl of Silicone oil. The samples were amplified for 40 cycles: denaturation at 95° for 45 sec, annealing at 52° for 1 min, and polymerization at 72° for 3 min plus 15 sec cumulatively added to each cycle. For PCR amplification of the first 10 Slow strains we employed the external primers shown in Figure 1: SODL (5′ → 3′) CCGAATTCCTGGATTCGTTTTTTATTT and SODR (5′ → 3′) CCGAATTCGTCGAGCAACAAGTGATAT, flanked by *Eco*RI restriction sites located *ca.* 410 bp proximal and 230 bp distal of the coding region of the *Sod* gene in *D. melanogaster* (KWIATOWSKI *et al.* 1989b, 1991). Amplification of all Fast strains and some Slow strains was done with primers derived from conserved regions of *Sod* and a downstream unidentified gene (J. KWIATOWSKI, unpublished results) (Figure 1): the N primer (5′ → 3′) CCTCTAGAAATGGTGGTTAAAGCTGTNTGCGT is derived from the first exon of the *Sod* gene, including the first 23 nucleotides of the coding region; and the O primer (5′ → 3′) ACGGAAGTCTAGAAGGGCTTTTTGGGCTTTGCCACCTG, derived from the downstream open reading frame (Figure 1). Both primers included artificial *Xba*I restriction sites for cloning purposes.

After amplification, we ran 5 μl of each sample on 0.7% agarose gel. PCR reactions with single DNA bands were cleaned using 100 μl of chloroform and precipitated with 50 μl of 7.5 M NH_4Ac and 300 μl of 95% ethanol. Samples were then restriction-digested with *Xba*I or *Eco*RI enzymes and run on a 0.7% agarose gel. The bands were excised out of the gel and run over Spin Bind DNA Extraction Units (FMC) as directed by the manufacturer and ligated to appropriately cut and phosphotased pUC19 or pUC21 vectors, then transformed into competent DH5α cells. DNA was isolated from the clones by standard procedures (MANIATIS *et al.* 1982).

Double-stranded DNA templates were sequenced by the dideoxynucleotide chain-termination method (SANGER *et al.* 1977) as described earlier (KWIATOWSKI *et al.* 1992) using the Sequenase I Kit (U.S. Biochemical Corp.) and 3000 Ci/mmol [^{32}P]dATP (Amersham), and run on 4–6% acrylamide gels at 45–50° for 3.5–7 hr. In addition to the PCR and M13 standard sequencing primers we used the following primers (derived from conserved *Drosophila Sod* coding regions): (5′ → 3′) C, CTTGCTGAGCTCGTGTCCACCCTTGCCCAGATCATC; I, GACATGCAGCCATTGGTGTTGTC; IR, GACAACACCAAYGGCTGCATGTC; CR, CAAGGGTGGACACGAGCTGAGCAAG (see Figure 1).

Except for haplotype Fast K, the low-level error rate of incorporation by the Taq polymerase (KWIATOWSKI *et al.* 1991) has been eliminated in our procedures by obtaining from each chromosome two independent amplifications and sequences

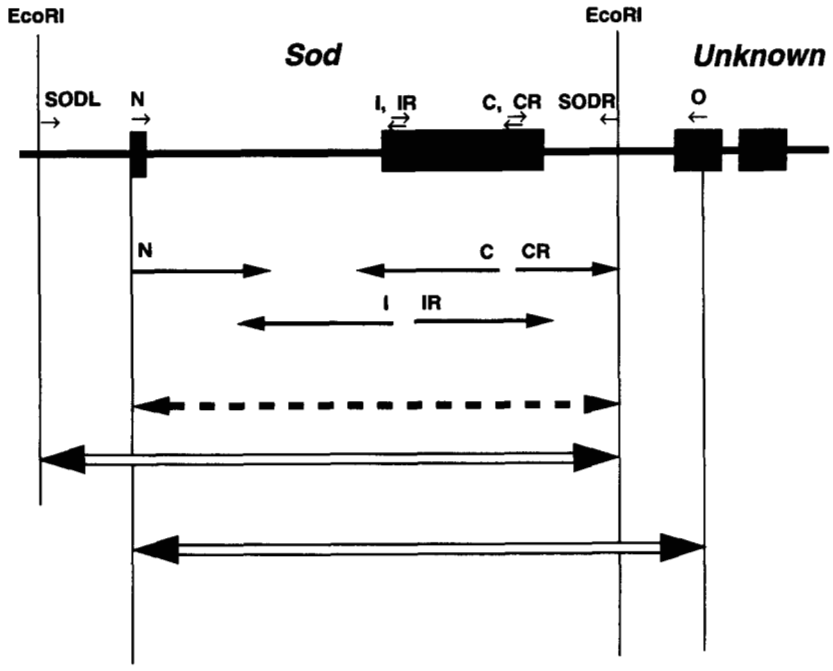

FIGURE 1.—Structure of the *Cu,Zn Sod* locus in *D. melanogaster* and strategies for PCR amplification, cloning and sequencing. The coding regions of the gene are shown as black boxes with primers shown above as short arrows. Primer N is located at the beginning of the first coding region while primers C/CR and I/IR are located 67 and 288 bp from the end of the second exon, respectively. (IR and CR are complementary to I and C from the opposite strand.) Primer O is located within the coding region of an adjacent downstream, unidentified gene. Primers SODL and SODR are located approximately 410 bp proximal and 230 bp distal of the *Sod* coding region, respectively. PCR amplification and outcome clones are represented by hollow arrows. The thinner arrows represent the direction and extent of sequencing from particular primers while the thick dashed arrow represents the region sequenced for this study.

to resolve any conflicts against our consensus *D. melanogaster Sod* sequence.

Statistical analysis: Certain parts of our statistical analysis require random samples from populations. Our samples are not random with respect to the number of Slow and Fast alleles. For example the Culver City sample contains, by design, 10 Slow alleles and 9 Fast alleles, while the frequency of Slow in the Culver City population is estimated to be 0.18 (see Table 1). The Slow alleles are, to the best of our knowledge, randomly chosen from among the Slow alleles in Culver City, and similarly the Fast alleles are randomly chosen Fast alleles. To construct a Culver City sample that can be treated at least approximately as a random sample for statistical analysis, we combined the 9 sampled Fast alleles with two Slow alleles. This artificial sample has the composition, in terms of Slow and Fast alleles, that would be most frequently observed in truly random samples from Culver City, given that the frequency of Slow is around 18%. We refer to this sample as the Culver City "constructed random sample" or, more briefly, the Culver City CRS. We treat this CRS as a random sample in the statistical tests and for estimation purposes. Although the P values of statistical tests with this sort of CRS have not been demonstrated to be the same as for truly random samples, we think that the p-values should be approximately the same and that estimates would have statistical properties very close to those based on random samples. The Barcelona CRS consists of one Slow allele and the nine sampled Fast alleles, which is consistent with our estimate of the frequency of the Slow allele in Barcelona of 0.076. A "Total CRS" is also considered, which consists of all 22 Fast alleles sequenced plus 3 Slow alleles.

TAJIMA's (1989) test is applied to assess whether the frequency spectrum of variation is compatible with neutrality. Also the test of FU and LI 1993) is applied. The levels of polymorphism and divergence between species are tested for compatibility with a simple neutral model by the method of HUDSON *et al.* (1987). This test will be referred to as the HKA test. To test for geographic differentiation the method of ROFF and BENTZEN (1989), which is based on haplotype frequencies, was used.

Recombination rates were estimated using the method of Hudson (1987). Nucleotide sites with 3 nucleotides or length polymorphisms were removed for estimating the recombination parameter.

A new test of neutrality based on the occurrence of subsets of the sample with low levels of variation was also applied. This test is described in DISCUSSION.

RESULTS

Protein electrophoresis: The results of the electrophoretic survey are shown in Table 1. The frequency of the Slow allele is in every case in the range 0.05 to 0.18.The estimated frequency of Slow in El Rio changed from 0.127 in October 1984 to 0.051 in October 1991. Statistically, the difference between these estimates is highly significant ($P < 0.001$ by Fisher's exact test). Similarly, the 1988 estimate of 0.175 in Culver City is significantly different from the October 1990 estimate of 0.076 for the Barcelona population ($P < 0.001$ by Fisher's exact test). Evidently, there are changes in the frequency of Fast and Slow alleles between years and also between localities.

Sequence analysis: As stated earlier, a total of 41 sequences are included in our analysis. The variation found in the samples is summarized in Figure 2 and Tables 2, 3 and 4. A total of 19 Slow allele sequences have been obtained, 10 from Culver City, 5 from Barcelona, 3 from El Rio and 1 from Davis. All 19 Slow alleles are identical in sequence over the 1410 bp examined. This sequence is shown aligned with a *D. simulans* sequence in Figure 3 in the APPENDIX. Among the 22 Fast alleles sequenced, there are ten distinct sequences, or haplotypes, designated Fast A through Fast E and Fast G through Fast K. The differences between these haplotypes and the Slow sequence are shown in Figure 2. The numbers of base pair differences between all pairs of haplotypes are shown in Table 2. All changes among the Fast alleles are synonymous. The Slow

TABLE 1

Estimated frequencies of the Slow allele

Locality	Date of collection	Genomes sampled	Frequency of Slow
El Rio	Oct. 1983	548	0.142
El Rio	Oct. 1984	448	0.127
El Rio	Oct. 1991	608	0.051
Culver City	Summer 1988	642	0.175
Barcelona, Spain	Oct. 1990	144	0.076

The sequenced Sod alleles from El Rio are all from the October 1991 sample.

sequence differs from the Fast A haplotype at a single site, site number 1013, which results in the single amino acid difference between Slow and Fast alleles. Table 3 shows the number of times that each haplotype was found in each locality. Summarizing: Fast A was found nine times, twice in Culver City, five times in Barcelona, once in El Rio and the Canton-S sequence is also Fast A. The Fast B was found twice, both times in Culver City. The Fast C was found four times, all in Culver City. The remaining seven haplotypes were each found just once in our sample of 22 Fast sequences.

There are a total of six insertion/deletion polymorphisms in the sample. From comparison with the *Sod* sequence of *D. simulans* (KWIATOWSKI *et al.* 1989), we can infer that four of these polymorphisms are due to deletions and one is due to an insertion; no inference about the sixth can be made. There is an interesting association of nucleotide site polymorphism with the insertion/deletion polymorphisms. In five out of six insertion/deletion polymorphisms, there is a nucleotide site polymorphism at or adjacent to the positions of the length polymorphisms.

It is important to consider whether there is evidence for differentiation of the populations in the different localities. It has already been demonstrated that the populations differ between years in the frequency of Fast and Slow alleles. This differentiation may be due to selection acting on the protein polymorphism, even in the face of strong migration. It is therefore of interest to assess to what extent the populations are differentiated with respect to the Fast sequences. To address this question we employ the test suggested by ROFF and BENTZEN (1989), using haplotype frequencies of Fast alleles in Culver City and Barcelona. A χ^2 statistic is calculated and its P value is estimated by 10,000 random permutations of the samples into two subsets, each of size 9. The observed χ^2 statistic is 12.3, and the estimated P value is 0.026. Thus, even with samples of only 9 Fast alleles from each locality, there is evidence of differentiation between the Barcelona population of Fast alleles in 1990 and the Culver City population of Fast alleles in 1988. Since this test is based only on Fast alleles, the differentiation is clearly not due to selection on the Fast/Slow polymorphism. The extent to which this differentiation of Barcelona and Culver City represents simply the action of isolation and drift, as opposed to natural selection is not clear. It is consistent with evidence from other studies that have shown differentiation between *D. melanogaster* populations on different continents (*e.g.*, HALE and SINGH 1991). The fact that several Fast haplotypes are multiply represented in the Culver City sample suggests the possibility that this population has recently experienced a bottleneck.

There are a total of 63 nucleotide site polymorphisms in the complete sample. Fifty-five sites are polymorphic in the Barcelona sample of sequences, 30 sites are polymorphic in the Culver City sample. At two sites, three different nucleotides were found segregating in our sample of sequences. Estimates of nucleotide diversity, π, (NEI 1987) from our samples are shown in Table 4. The nucleotide diversity for the entire 1410-bp region sequenced is about 0.009 in the Barcelona CRS, the Culver City CRS and in the Total CRS. (See the MATERIALS AND METHODS for a definition of CRS.) The intron and the silent sites have a π of about 0.013. The average number of differences per base pair between a sequence from Culver City and a sequence from Barcelona is 0.010. The nucleotide diversity values at *Sod* are compared to the nucleotide diversity estimates from several other loci in Table 5. The nucleotide diversity levels for *Sod* are approximately in the middle of the rather large range of values observed at other loci. Also included in the table are estimates of $\theta = 4Nu$, based on the number of segregating sites and assuming an equilibrium infinite-sites neutral model (WATTERSON 1975). (N is the effective population size and u is the neutral mutation rate per base pair.) Note that under the neutral model, π is also an estimate of θ. (For sex-linked loci, estimates of θ are slightly different.)

The frequency spectrum of the variation found in the *Sod* region can be tested for compatibility with the neutral model with TAJIMA's (1989) test, which is based on comparing π and the estimate of θ. The values of TAJIMA's D statistics are 1.02 and −1.32 for the Culver City CRS and the Barcelona CRS, respectively. Neither of these values is significantly different from zero, and thus provides no grounds for rejecting the neutral model. The test of FU and LI (1993) gives similar results.

The number of differences per base pair between the Slow sequence and the available *D. simulans* sequence (KWIATOWSKI *et al.* 1989a) is 0.096 for silent sites, and 0.071 for intron sites. As can be seen from Table 5, the amount of divergence between *D. melanogaster* and *Drosophila simulans* at intron sites is very consistent among five loci for which data are available, ranging from 0.047 for *Adh* to 0.071 for *Sod*. For silent sites of the exons, the variability in divergence is greater, ranging from 0.052 for *Adh* to 0.15 for *Pgd*. There are several loci, however that cluster at a divergence of about 0.1. *Sod* falls in this group of loci. Thus at both silent sites and intron sites, *Sod* is fairly typical in terms of levels of polymorphism and divergence. The HKA

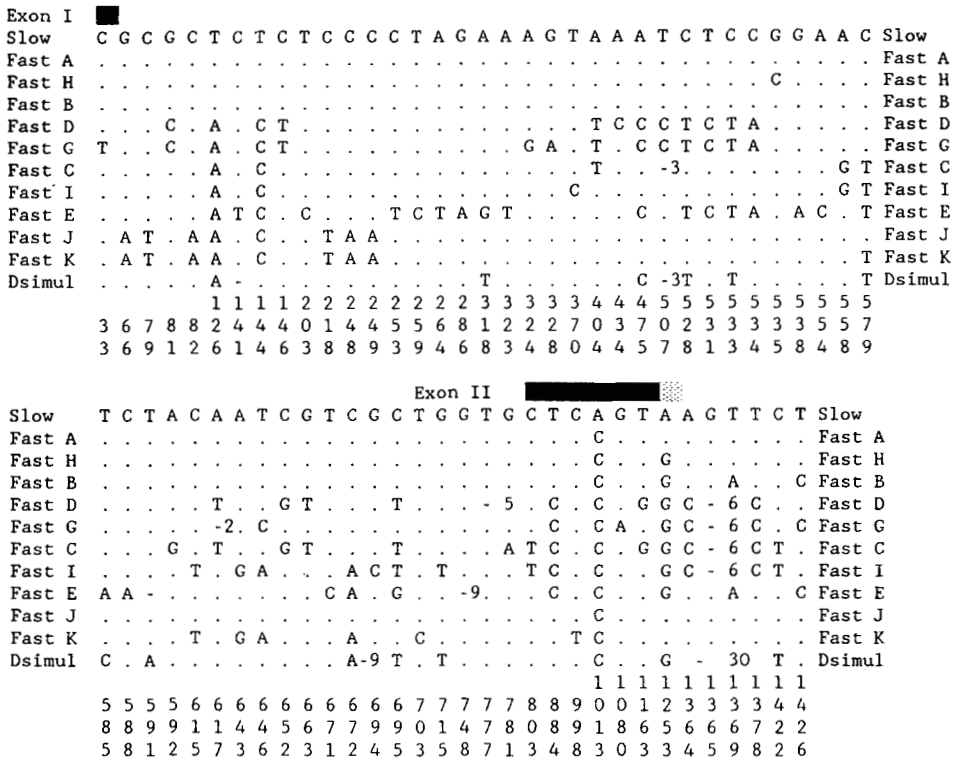

FIGURE 2.—Polymorphic sites in *D. melanogaster Cu,Zn Sod*. Exon regions are in black (shaded if not translated.) The sites are numbered at the bottom, using the numbering of Figure 3 in the APPENDIX, in which position 1 is immediately after the PCR primer N (position 22 of the coding region of the first exon.) For those sites which are polymorphic within *D. melanogaster*, the *D. simulans* nucleotides are given in the row labeled Dsimul. Length polymorphisms are indicated by "–", with the number of the first inserted/deleted base indicated at the bottom. If the insertion/deletion is longer than one base pair, the length is indicated after the "– ."

TABLE 2

Pairwise differences between haplotypes

Haplotype	Fast A	Fast H	Fast B	Fast J	Fast K	Fast I	Fast C	Fast D	Fast G	Fast E
Slow	1	3	4	9	16	19	19	22	22	28
Fast A		2	3	8	15	18	18	21	21	27
Fast H			3	10	17	18	18	21	21	27
Fast B				11	18	18	18	21	19	24
Fast J					7	22	22	25	25	31
Fast K						19	27	32	31	34
Fast I							14	25	26	31
Fast C								13	23	33
Fast D									11	28
Fast G										27

test (HUDSON *et al.* 1987) compares the levels of polymorphism and divergence between two or more loci, and assesses the compatibility of these values with the neutral model. Applying this test to the Barcelona *Sod* data, and using the 5′-flanking region of *Adh* as a second locus, yields a χ^2 value of 1.2 (1 d.f.). This indicates that the levels of polymorphism and divergence at the *Sod* locus are compatible with neutrality, assuming that the Adh-5′ region provides a valid neutral standard. Similarly, the Culver City CRS, and Total CRS show no significant departures from neutrality, as assessed by the HKA test.

The sequence data clearly indicate that the Slow/Fast polymorphism is not an old balanced polymorphism. The Slow alleles show no variation among themselves, even when sampled from localities as distant as California and Spain. The Slow allele sequence differs from the Fast A haplotype at only a single nucleotide site, the site which produces the amino acid difference between them. In the sample of Fast alleles as a whole, Fast A is the most common haplotype, constituting 9 out of 22 Fast sequences sampled. It is also very common in each of the localities: Culver City, El Rio, and Barcelona. (See Table 3.) The Canton S sequence is also Fast A. The Fast H haplotype and Fast B haplotype are very similar to the Fast A haplotype, differing from it by two and three nucleotides respectively. The *D. simulans* SOD enzyme is identical in amino acid sequence to the Fast allele. These data clearly indicate that the Slow allele arose quite recently from the Fast A haplotype. The large number of sequences (mostly Fast A alleles) which are so similar to each other suggests that selection has increased the frequency of some haplotypes. This issue will be addressed in the DISCUSSION.

Making inferences about rates of recombination (or conversion) from polymorphism data is difficult. Available methods of estimating recombination rates

TABLE 3

Distribution of haplotypes in samples from different localities

Haplotype	Locality/Stock					
	Culver City	Barcelona	El Rio	Oregon R	Canton S	Davis
Slow	10	5	3			1
Fast A	2	5	1		1	
B	2					
C	4					
D	1					
E		1				
G		1				
H				1		
I		1				
J			1			
K		1				

(HUDSON and KAPLAN 1985; HUDSON 1987) from this sort of data rely on the assumption that an equilibrium neutral model is appropriate, whereas we will argue later that our samples are incompatible with the neutral model. Nevertheless these methods may provide some information and so we present them here.

First, we note that recombination is suggested by the pattern of similarity and differences between the haplotypes. For example, the Fast J haplotype differs from Fast A at eight sites in the first 250 base pairs of the region sequenced (starting at the 3′ end), but is identical to Fast A in the remaining 1160 bp. This suggests that the gene tree for the 3′ end of the region is different from the gene tree of the 5′ end of the region, which is only possible if recombination has occurred in the history of our sample. Similarly, Fast B is identical to Slow in positions 1 to 1012, and differs from Slow at four sites between position 1013 and 1428, whereas Fast E differs from Slow at 24 sites in the first 1012 sites, but is the same as Fast B in positions 1013 to 1428. This clustering of similarities and differences suggests the occurrence of recombination in the history of the sample (STEPHENS 1985; SAWYER 1989; HARTL and SAWYER 1991). Using the four-gamete rule of HUDSON and KAPLAN (1985) five recombination events can be inferred from the data. (These inferred recombination events could also be explained by multiple hits at single nucleotide sites.) These five inferred recombination events occurred in the intervals 534–579, 579–617, 672–695, 803–1163 and 1378–1426, where the numbers refer to nucleotide positions numbered as in Figure 3. The number of recombination events in this region of *Sod* is similar to the number of events inferred in the history of a sample of *Adh* alleles of comparable size (KREITMAN 1983; HUDSON and KAPLAN 1985). This suggests that the rate of recombination in the *Sod* region may be roughly the same as in the *Adh* region.

An estimate of Nr, where N is the effective population size and r is the recombination rate between adjacent base pairs per generation, can be obtained by the method of Hudson (1987) assuming an equilibrium neutral model. Applying the method to the Total CRS we obtain an estimate of Nr of 0.0004. For the Barcelona CRS the estimate is also about 0.0004. These estimates seem low in light of estimates of r from other loci in *D. melanogaster*. For example, if we take $r = 10^{-8}$, as has been estimated for the *rosy* locus of *D. melanogaster* (CHOVNICK *et al.* 1977) and $N = 10^6$, then the product, Nr, is 0.01. It may be that this low estimate of Nr is a result of recent selection in the *Sod* region, which as we will argue later, resulted in the high frequency of Fast A. Prior to this selection the frequency of Fast A would have been much lower. If we consider a modified Barcelona sample with four Fast A's removed, leaving just one Fast A allele plus the five other sampled alleles, then the estimated value of Nr is 0.011, which is very similar to the value obtained with the direct estimates of r. Even under the equilibrium neutral model these estimates have high variances and given that the equilibrium neutral model is unlikely to be appropriate for this locus, these estimates should not be given too much weight. We conclude only that no strong case can be made that recombination rates are extraordinarily low or high at *Sod*.

TABLE 4

Levels of polymorphism at *Sod* in *D. melanogaster*

Site	Entire sequence (1410 bp)		Silent sites (109 sites) [a]		Intron (706 bp)	
	π	S	π	S	π	S
Barcelona CRS (n = 10)	0.010	55	0.012	5	0.016	43
Culver City CRS (n = 11)	0.008	30	0.015	3	0.010	20
Total CRS (n = 25)	0.009	63	0.014	5	0.013	51

π is nucleotide diversity. S is the number of polymorphic sites.
[a] 109 is the number of silent site equivalents (KREITMAN 1983) of the coding region sequenced.

DISCUSSION

Our samples have revealed levels of variation that are fairly typical of other regions of the *D. melanogaster* ge-

TABLE 5

Estimates of divergence between *D. melanogaster* and *D. simulans*, θ, and nucleotide diversity in *D. melanogaster*

Locus	Divergence[a]	θ	π	Divergence/π	bp	Reference
Silent sites						
Adh	0.052	0.023	0.0286	1.8	192	KREITMAN and HUDSON (1991)
Adh-dup	0.13	0.009	0.005	26.0	181	KREITMAN and HUDSON (1991)
Gpdh	0.062	0.028		2.2[b]	251	TAKANO *et al.* (1993)
ci	0.11	0.0	0.0	∞	218	BERRY *et al.* (1991)
Sod	0.096	0.016	0.0093	10.0	109	This study (Barcelona CRS)
G6pd	0.11	0.019[c]	0.015	5.8[d]	410	EANES *et al.* (1993)
Pgd	0.15	0.0038[c]		39.0[e]	1420	BEGUN and AQUADRO (1993)
Introns						
Adh	0.047	0.018	0.017	2.8	127	KREITMAN and HUDSON (1991)
Adh-dup	0.062	0.007	0.0063	9.8	419	KREITMAN and HUDSON (1991)
Gpdh	0.051	0.0095		5.4	2454	TAKANO *et al.* (1993)
Sod	0.071	0.021	0.016	4.6	706	This study (Barcelona CRS)
Pgd	0.060	0.0024[c]		25.0[e]	336	BEGUN and AQUADRO (1993)
5′-Flanking region						
Adh-5′	0.062	0.008	0.009	6.9	1243	KREITMAN and HUDSON (1991)

[a] Divergence is the number of differences between species divided by the number of sites.
[b] This value is actually divergence divided by the estimate of θ.
[c] These are *X*-linked loci. θ is estimated by multiplying the usual estimate by 4/3.
[d] Since *G6pd* is *X*-linked, the number here is 0.25 + divergence/$(4\pi/3)$.
[e] Since *Pgd* is *X*-linked, and only an estimate of θ is available, the number here is 0.25 + divergence/θ.

nome. (See Table 5.) There are a total of 63 nucleotide polymorphisms in our set of 41 sequences. Remarkably, however, a subset of 31 of these sequences have only five sites that are polymorphic among them. This subset of 31 includes 19 Slow allele sequences which are absolutely identical to each other. Since the Slow allele is at a frequency of around 10% in these populations, a sample of these alleles under neutrality is expected to harbor only about 10% as much variation as a random sample (HUDSON and KAPLAN 1986). Even taking this into account, there is a large subset of sequences, consisting mostly of Fast A alleles, which show very little variation. For instance, if we consider our Total CRS, which is a "constructed random sample" consisting of all our Fast allele sequences plus 3 Slow allele sequences, for a total of 25 sequences, there is a subset of 12 sequences (three Slow sequences and 9 Fast A sequences)with just one polymorphic site. Consider also the Barcelona CRS, which has one Slow sequence and nine Fast sequences, including five Fast A sequences. This CRS has a subset of five sequences with no variation among them, and a subset of six sequences with only one site polymorphic. This pattern of variation suggests that a rare variant (perhaps a new mutation) has risen rapidly in frequency due to natural selection. As it increased in frequency, the haplotype in which it is embedded was pulled up in frequency at the same time. It should be emphasized that the large subset of sequences with little variation is mostly, or in some cases wholly, made up of Fast A alleles with the Slow alleles making up a very small part of the subset. It appears that a selection hypothesis to explain our data must involve more than selection distinguishing Fast from Slow, because such selection would not explain the high frequency of a particular Fast haplotype. It should also be emphasized that the variation upon which selection is acting is not necessarily within the region that we have sequenced, but it must be tightly linked to it. Before giving further consideration to selection hypotheses, it behooves us to test the neutral null hypothesis, to make sure that this pattern in our data is indeed very unlikely under a simple model with drift and mutation. We now propose a statistical procedure for carrying out such a test and we apply it to our data.

The general idea of our test as follows. Suppose that we have a sample of n sequences, with m polymorphic sites. Suppose also that there is a subset consisting of i of these sequences that has only j sites polymorphic. We wish to know whether this sample is compatible with neutrality, *i.e.*, whether the probability of finding such a subset with j or fewer polymorphisms is too small under the null hypothesis of neutrality for us to accept neutrality? Our test is carried out by estimating the probability that, in a sample of size n, there is a subset of size i with j or fewer polymorphic sites in it, given that the sample as a whole has m polymorphic sites.

To estimate this probability many computer generated random samples of size n are produced, each with m polymorphic sites, under a Wright-Fisher equilibrium neutral model. These samples are produced using a computer algorithm based on the coalescent process (HUDSON 1983, 1993). To generate each sample, a random genealogy is produced and m mutations are distributed onto the genealogy. (Note that the number of mutations, m, is a constant in these simulations and that the usual mutation parameter, $4Nu$, does not appear as a parameter. The rationale for this type of simulation has been discussed elsewhere (HUDSON 1993).) These simulations can be carried out with any specified level of recombination. Each sample was then examined by computer to determine if indeed there was a subset of size

i which had j or fewer polymorphic sites in the subset. The fraction of samples which contained such a subset is an estimate of the P value of the observed sample. In all the following applications of this test, the P value is estimated from 10,000 computer generated samples.

There are several different samples and different subsets to which we could apply our test. The Barcelona CRS consists of 10 sequences with 55 polymorphic sites. There is a subset of this sample of size five with no polymorphisms, and a subset of six with one polymorphism. With no recombination, the estimated P value of the sample with subset of size five with no polymorphism is 0.011. The estimated P value for a subset of size six with one polymorphism is 0.007. With recombination, the probability of such homogeneous subsamples decreases. Under the neutral model, the properties of such samples depends on the product of effective population size (N) and the recombination rate between adjacent base pairs per generation (r). As described in RESULTS, a rough estimate of Nr is 0.01, based on estimates of r from the *rosy* locus and assuming that N is 10^6. With this level of recombination the estimated P values of the Barcelona CRS are 0.0 (zero out of 10,000 samples) for the subsample of size five and the subsample of size six. Even with a level of recombination one-tenth of this, the estimated P values are still only 0.005 and 0.003, for the subsample of size five and six, respectively.

We can also apply this test to a CRS consisting of all 22 Fast alleles and three Slow alleles, which have a total of 63 polymorphic sites. The estimated P value of such a sample with a subset of 12 sequences with just one polymorphic site is 0.0081, 0.0030 and 0.0 with $Nr = 0$, 0.001 and 0.01, respectively.

Recall that the Culver City CRS has two Fast A alleles which together with the two Slow alleles form a subset of size four with just 1 polymorphic site. The Culver City sample as a whole has 30 polymorphic sites. This subset of four in a sample of 11 is not a significant departure from neutrality for $Nr = 0$. But if Nr is greater than 0.005 then the estimated P value is less than 0.05. For $Nr = 0.01$, the estimated P value for such a subset is 0.003. Note that the Fast C allele is present four times in the Culver City CRS. Such a subset, four alleles with no polymorphism is also quite unlikely under the equilibrium neutral model when $Nr = 0.01$. Thus, if recombination rates for the *Sod* region are this large, the Culver City sample is very unusual (under the equilibrium model) with respect to two different subsets, the Slow/Fast-A subset and the Fast C subset.

Our statistical test for high frequency haplotypes was designed and applied after examining the data and noting what appeared to us to be an unusual pattern, namely a group of haplotypes at high frequency with little variation among them. In other words our test is an *a posteriori* test. The outcome of our test as applied to the *Sod* data should be interpreted with that in mind. It is not clear how to do it, but perhaps the P values should be adjusted in some way to reflect the fact that the data were examined for several characteristics, any one of which might have been used to reject the neutral null hypothesis. We have in some sense applied multiple tests, as least in some informal way, and as a consequence perhaps something like a Bonferroni correction for multiple tests should be applied. It is not clear to us how to make such an adjustment. Nevertheless, it is our view that the estimated P value for the Barcelona sample is so low ($P < 0.0001$, for the most plausible recombination rate) that any such considerations would not alter the basic conclusion that the Barcelona sample is not compatible with the equilibrium neutral model.

How might we account for the very unusual pattern of variation shown by the Barcelona sample? First we consider some neutral alternatives. The test assumes an equilibrium infinite-sites neutral model with panmixis. Mutations are assumed to occur independently at different sites. It is known that many mutational mechanisms can produce changes at several sites simultaneously (*e.g.*, GOLDING 1985). These mutational mechanisms would tend to produce departures from our null model in the direction that we have observed, that is with relatively few haplotypes for the number of polymorphic sites observed. However, to explain our data, such mutational mechanisms would have to account for a large number of the observed polymorphisms. This is because even if the number of polymorphic sites in the Barcelona sample were only 35 (instead of the 55 that was observed), the estimated P value of a subsample of size 6 having only one polymorphic site is 0.029 for the case of $Nr = 0.0$. For higher recombination rates the estimated P value is much lower. It seems unlikely that complex mutational mechanisms can account for 20 or more mutations in our sample, and thus this mechanism is not an adequate explanation for the Barcelona sample.

The null hypothesis for our test is a neutral model without geographic subdivision, while we have already shown that some differentiation has occurred between the Barcelona population and the Culver City population. With a strongly subdivided population, there will certainly be higher probabilities of subsets with little variation within them. This is because with strong subdivision, much of the variation will be variation which distinguishes subpopulations, and so subsets consisting of sequences from within subpopulations are more likely to be very similar for any given overall level of variation. In addition, STROBECK (1987) has shown that the number of haplotypes within subpopulations will tend to be low for a given level of nucleotide diversity, compared to panmictic models. However, in our case the subset of homogeneous sequences is made up of haplotypes which are the most common haplotypes across all localities examined. It seems likely that this pattern is even more improbable in subdivided populations than in a panmictic population.

Can the data be explained by a neutral model with a history of bottlenecks and population expansion, together with subdivision? Changes in population size can certainly have a dramatic effect on the sampling properties under the neutral model (TAJIMA 1993). At this time, we cannot rule out the possibility that recent population size changes (bottlenecks and expansions) with or without geographic structure have played a major role in generating the unusual pattern we have observed. Such alternatives warrant further investigation. However, it is important to keep in mind that any hypothesis must explain the high frequency of Fast A (together with Slow) in geographically widely separated populations, including the Barcelona population, the El Rio population and the Culver City population. Also, a history of population size changes will have affected all autosomal loci, and therefore hypotheses of this type have implications for all autosomal loci in *D. melanogaster*. For any such hypothesis to be tenable it must be consistent with the patterns of variation observed at other loci in samples from North America and Europe. The Culver City sample, which exhibits high frequency of the Fast C, the Fast B and the Fast A (plus Slow), may well reflect a bottleneck in the recent history of the Culver City population.

We now consider some alternative hypotheses involving selection. Recall that the Fast A haplotype constitutes 50% of the haplotypes in the Barcelona CRS, while the Slow allele is estimated to be at a frequency of about 8% in Barcelona. One scenario consistent with the data is the following. A mutation arose in the *Sod* locus or at a tightly linked locus and its frequency rapidly rose owing to selection to about 40%. The Fast A haplotype rose to frequency of 40% by hitchhiking along with this rare variant. Then the amino acid mutation, Fast → Slow, arose in a Fast A haplotype. Subsequently, either because of selection on the amino acid variant itself, or simply by hitchhiking along with the first mutation (which continued to rise to about 50%), the frequency of the Slow allele rose from essentially zero to 5–18%, depending on the locality. Notice that this scenario involves selection that first increases the frequency of Fast A haplotypes, relative to other Fast haplotypes, and secondarily may include selection on the Slow allele. If selection had acted only on the amino acid change, the high frequency of Fast A could not be produced.

The size of the region that would hitchhike along with such a selected variant could be quite large and would depend on the strength of selection and the rate of recombination in this region. The analysis of hitchhiking models in which mutations sweep to fixation, gives some insight into the size of such regions. KAPLAN *et al.* (1989) show that a mutant with selective advantage of 0.001 that sweeps through a population will reduce variation at sites up to 1000 bp away from the site, assuming typical levels of *D. melanogaster* recombination. Stronger selection could clearly sweep out much larger regions. By examining DNA sequence variation at varying distances from the *Sod* locus one could delimit the region that exhibits the non-neutral pattern detected at the *Sod* locus, and thus delimit the region within which the selected mutation or mutations must reside. Since the strength of selection is an important parameter, along with the recombination rate, in determining the size of the region affected by a selected mutation, information on the size of the region affected would also permit one to estimate the strength of selection acting on the selected mutant.

What form of selection is needed to account for the pattern of variation that we see? First, we emphasize once again that the selection we postulate could be acting on variation in the *Sod* region sequenced or may be acting at a tightly linked site outside the region sequenced. The variant which is increasing in frequency may actually be favored over all existing variants and thus on its way to fixation or it may be a new balanced polymorphism. If it is on its way to fixation, the observed pattern is a very ephemeral one that we have by chance caught when the new variant is at a frequency of about one-half. In this case, in a fairly short time the sweeping variant will reach a frequency of one, and most variation in the region surrounding the favored mutation will be eliminated. Another possibility is that the new variant represents a new balanced polymorphism. In this case, the rare variant has recently risen to intermediate frequency, where it may now be at or near some sort of equilibrium. The equilibrium may be different in different localities and may be temporally varying as well. This would account for the different frequency of Slow in different years and different localities. The form of the natural selection could be any of the diversity preserving forms of selection such as frequency dependent selection, heterozygote superiority, certain forms of temporal and spatial fluctuations of selection coefficients, as well as spatial variation in selection coefficients with limited gene flow.

The rejection of neutrality, using our test based on the amount of variation in a subset of sequences, contrasts dramatically with the results of the Tajima test and the HKA test. These results suggest that these two commonly applied tests may not be very powerful against many alternative hypotheses, especially when the data sets are small. In fact, under the selective scenario that we have suggested, the HKA test and the Tajima test are unlikely to detect departures from neutrality. To see this, consider how π and S, the number of polymorphic sites in a sample, are likely to be affected by the increase in frequency of a random haplotype. To be concrete, suppose that a random haplotype in a population at equilibrium under the neutral model, instantaneously increases in frequency to 50%. This will decrease π by 25% (the probability that two randomly chosen copies are both of the newly increased haplotype.) How is S affected by our putative selection event? A sample of size

10 after the selective increase of one haplotype, is essentially like a sample of size six taken before selection which is then modified by picking one haplotype and increasing its representation in the sample to five copies, to make up a total sample of size 10. The process of increasing the representation of one haplotypte to five copies has no effect on the number of polymorphic sites, so the number of polymorphic sites in the post selection sample of size 10 is expected to be approximately like the number in a sample of size six before the selection. The expected number of polymorphic sites in a sample under the neutral model is 2.83θ for a sample of size 10 and 2.28θ for a sample of size 6 (WATTERSON 1975), or a decrease of 19%. Thus, π and S are affected by about the same amount, and hence TAJIMA's test, which is based on comparing π and S, is unlikely to detect departures from neutrality under our selection model. The HKA test is based on levels of polymorphism and divergence. The selection that we propose would have essentially no effect on divergence between species, but would decrease π and S by about 25%. It would be useful to carry out an analysis of the power of the HKA test to detect a level of polymorphism 25% below the expected level. In the absence of such an analysis, it is difficult to be sure, but it may well be that such departures are unlikely to be detected with the sample size that we have utilized.

The results of the HKA test must be interpreted with caution for another reason as well. The HKA test contrasts the levels of polymorphism and divergence between two (or more) loci. If the neutral model is rejected, one cannot without additional information, attribute the rejection to one of the loci. Either of the loci, or both could be the cause of the rejection. Furthermore, if two loci both depart from neutrality in the same way, say by having an elevated level of polymorphism due to balancing selection, then the HKA test applied to these two loci will certainly fail to detect any departure from neutrality. It is possible to test a single locus for departure from neutrality using the HKA test, if one has data from a "neutral locus," for which the assumptions of the neutral model are believed to hold. The neutral locus can be used as a standard for comparison to other loci. The problem is that we do not have a clearly neutral locus to serve as a standard. This is in part because such a large fraction of the loci examined to date, exhibit patterns of DNA polymorphism that suggest the impact of natural selection.

To illustrate further the problem of finding a neutral locus to use as a standard in the HKA test, we consider the ratio of divergence between species to π for several loci. Under the infinite-site neutral model, the number of differences per base pair between species, is an estimate of $\theta(t/2N + f)$ (GILLESPIE and LANGLEY 1979) where t is the number of generations since speciation and f is the ratio of the effective population size of the ancestral population that gave rise to the two species and N. Under the neutral model this quantity is expected to vary from locus to locus, as mutation rates and levels of constraint vary between loci. However, since π estimates θ under neutrality, the ratio of divergence to π, can be considered an estimate of $(t/2N + f)$, which is the same for every locus. The HKA test is, in essence, a test of whether the ratio of divergence to π is sufficiently constant for a set of loci, to be consistent with the neutral model. (Actually the original version of the test used divergence and estimates of θ for the test.) The ratio of divergence to π is shown in Table 5, for several loci. A wide range of values for this ratio are evident in the table. This wide range of values is also seen in surveys of restriction site polymorphisms (BEGUN and AQUADRO 1992). Proper statistical tests of neutrality have been applied to many subsets of these data and have resulted in some cases in rejection of the neutral model and a variety of selective hypotheses have been put forward to explain these departures from neutrality. The problem for us now is: Are any of the loci listed in Table 5 appropriate for use as a standard in the HKA test? One might reasonably speculate that the true value of $(t/2N + f)$ is around 6 and that those loci with ratios of divergence to π, in the range of 4 to about 10, have variation that is best explained by the neutral model. If this speculation were correct, it would be appropriate to use, for example, the Adh 5′ region as a standard neutral locus for use in testing the DNA variation at each new locus that was surveyed. Although it seems reasonable to use Adh 5′ as a standard provisionally, there remains the possibility that Adh 5′ is very strongly influenced by selection and gives a poor estimate of $(t/2N + f)$. Given the small amount of data available at this point and the great variation in the ratio of divergence to π that has been observed, this possibility certainly cannot be ruled out.

The application of the HKA test has clearly shown that the data from the loci in Table 5 are not compatible with the simple version of the neutral model which is used as a null model. What is not clear is which loci are responsible for the departures from neutrality, and therefore which loci, if any, are appropriate for use as a neutral standard for testing additional loci.

This research was supported by U.S. Public Health Service grant GM42397.

LITERATURE CITED

AYALA, F. J., J. R. POWELL, M. L. TRACEY, C. MOURAO and S. PEREZ-SALAS, 1972 Enzyme variability in the *Drosophila willistoni* group. IV. Genic variation in natural populations of *Drosophila willistoni*. Genetics **70:** 113–139.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

BEGUN, D. J., and C. F. AQUADRO, 1993 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: selection and geographic differentiation. Genetics **136:** 155–171.

BERRY, A. J., J. W. AJIOKA and M. KREITMAN, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

CEBALLOS, I., F. JAVOY-AGID, A. DELACOURTE, A. NICOLE and P. M. SINET, 1990 Parkinson's disease and Alzheimer's disease: neurodegenerative disorders due to brain antioxidant system deficiency? Adv. Exp. Med. Biol. **264:** 493–498.

CEBALLOS-PICOT, I., A. NICOLE and P. M. SINET, 1992 Cellular clones and transgenic mice overexpressing copper-zinc superoxide dismutase: models for the study of free radical metabolism and aging. Exs **62:** 89–98.

CHOVNICK, A., W. GELBART and M. MCCARRON, 1977 Organization of the Rosy locus in *Drosophila melanogaster*. Cell **11:** 1–10.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the G6PD gene in the *Drosophila melanogaster* and *D. simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

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

GILLESPIE, J. H., and C. H. LANGLEY, 1979 Are evolutionary rates really variable? J. Mol. Evol. **13:** 27–34.

GOLDING, G. B., and B. W. GLICKMAN, 1985 Sequence-directed mutagenesis: evidence from a phylogenetic history of human alpha-interferon genes. Proc. Natl. Acad. Sci. USA **82:** 8577–8581.

GRAF, J. D., and F. J. AYALA, 1986 Genetic variation for superoxide dismutase level in *Drosophila melanogaster*. Biochem. Genet. **24:** 153–168.

HALE, L. R., and R. S. SINGH, 1991 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. IV. Mitochondrial DNA variation and the role of history *vs.* selection in the genetic structure of populations. Genetics **129:** 103–117.

HARTL, D. L., and S. A. SAWYER, 1991 Inference of selection and recombination from nucleotide sequence data. J. Evol. Biol. **4:** 519–532.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, Mass.

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., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

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., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **355:** 167–170.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KAWASAKI, E. S., 1990 pp. 146–152 in *PCR Protocols: A Guide to Methods and Applications*, edited by M. A. INNIS, D. H. GELFAND, J. J. SNINSKI and T. J. WHITE. Academic Press, San Diego.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

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

KWIATOWSKI, J., F. GONZALEZ and F. J. AYALA, 1989a *Drosophila simulans* Cu-Zn superoxide dismutase gene sequence. Nucleic Acids Res. **17:** 6735.

KWIATOWSKI, J., M. PATEL and F. J. AYALA, 1989b *Drosophila melanogaster* Cu,Zn superoxide dismutase gene sequence. Nucleic Acids Res. **17:** 1264.

KWIATOWSKI, J., D. SKARECKY, S. HERNANDEZ, D. PHAM, F. QUIJAS *et al.*, 1991 High fidelity of the polymerase chain reaction. Mol. Biol. Evol. **8:** 884–887.

KWIATOWSKI, J., D. SKARECKY and F. J. AYALA, 1992 Structure and sequence of the Cu,Zn SOD gene in the Mediterranean fruit-fly *Ceratitis capitata*: intron insertion and deletion in the evolution of the SOD gene. Mol. Phyl. Evol. **1:** 72–82.

LEE, Y. M., and F. J. AYALA, 1985 Superoxide dismutase in *Drosophila melanogaster*. Mutation site difference between two electromorphs. FEBS Lett. **179:** 115–119.

LEE, Y. M., D. J. FRIEDMAN and F. J. AYALA, 1985 Complete amino acid sequence of copper-zinc superoxide dismutase from *Drosophila melanogaster*. Arch. Biochem. Biophys. **241:** 577–589.

LEE, Y. M., H. P. MISRA and F. J. AYALA, 1981 Superoxide dismutase in *Drosophila melanogaster:* Biochemical and structural characteristics of allozyme variants. Proc. Natl. Acad. Sci. USA **78:** 7052–7055.

MANIATIS, T., E. F. FRITSCH and J. SAMBROOK, 1982 *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

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

PENG, T., A. MOYA and F. J. AYALA, 1986 Irradiation-resistance conferred by superoxide dismutase: Possible adaptive role of a natural polymorphism in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **83:** 684–687.

PENG, T. X., A. MOYA and F. J. AYALA, 1991 Two modes of balancing selection in *Drosophila melanogaster*: overcompensation and overdominance. Genetics **128:** 381–391.

ROFF, D. A., and P. BENTZEN, 1989 The statistical analysis of mitochondrial DNA polymorphisms: chi-square and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

ROSEN, D. R., T. SIDDIQUE, D. PATTERSON, D. A. FIGLEWICZ, P. SAPP *et al.*, 1993 Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature **362:** 59–62.

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

SANGER, F., S. MIKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SAWYER, S. A., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

SEAGER, R. D., and F. J. AYALA, 1982 Chromosome interactions in *Drosophila melanogaster*. I. Viability studies. Genetics **102:** 467–483.

SETO, N. O., S. HAYASHI and G. M. TENER, 1989 Cloning, sequence analysis and chromosomal localization of the Cu-Zn superoxide dismutase gene of *Drosophila melanogaster*. Gene **75:** 85–92.

SINGH, R. S., D. A. HICKEY and J. DAVID, 1982 Genetic differentiation between geographically distant populations of *Drosophila melanogaster*. Genetics **101:** 235–256.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

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

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

TAJIMA, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, Mass.

TAKANO, T. S., S. KUSAKABE and T. MUKAI, 1993 DNA polymorphisms and the origin of protein polymorphism at the Gpdh locus of *Drosophila melanogaster*, pp. 179–190 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, Mass.

TANIGUCHI, N., 1992 Clinical significances of superoxide dismutases: changes in aging, diabetes, ischemia, and cancer. Adv. Clin. Chem. **29:** 1–59.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **10:** 256–276.

Communicating editor: A. G. CLARK

APPENDIX

Alignment of sequences is shown in Figure 3 (p. 1340).

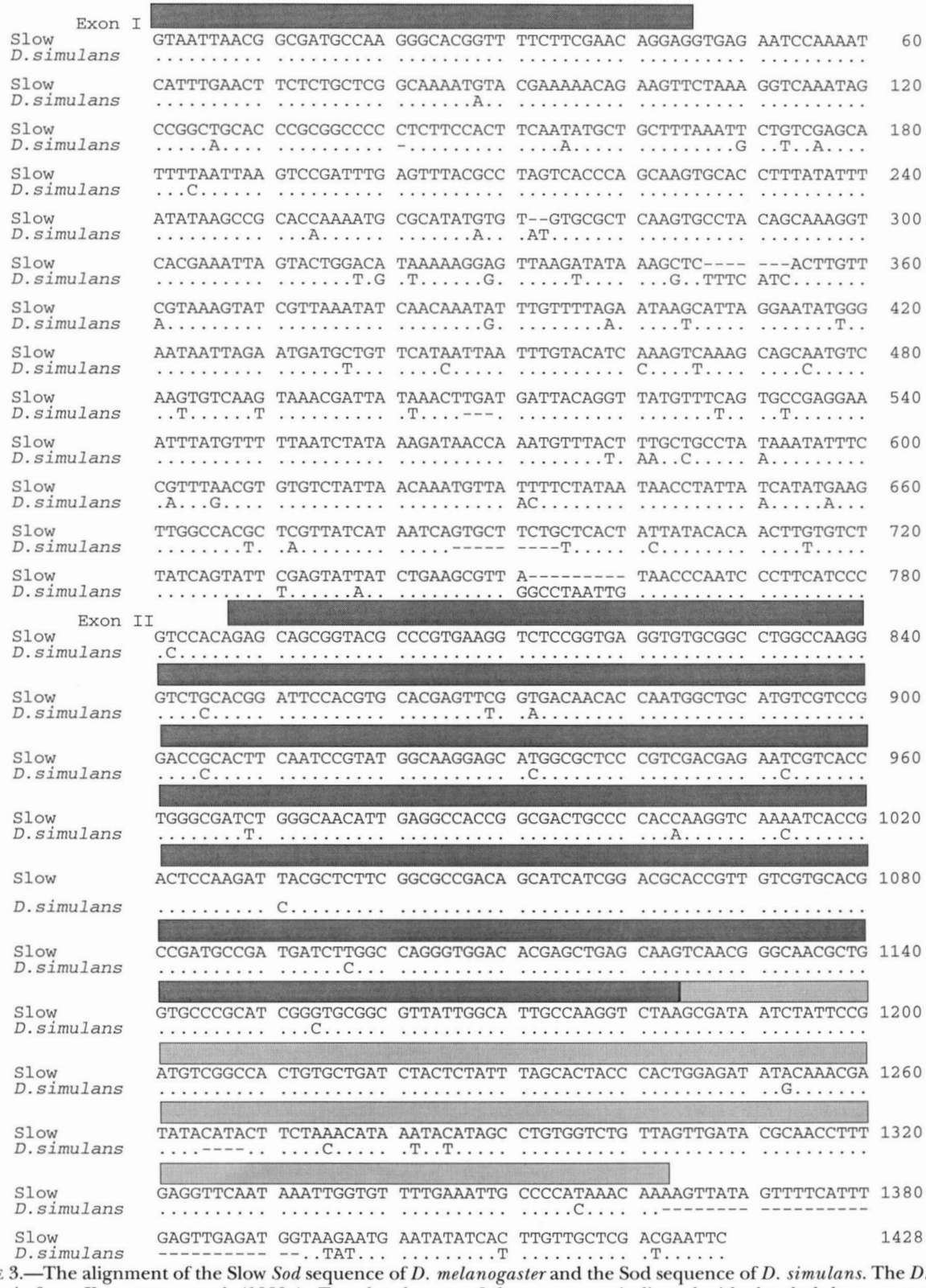

FIGURE 3.—The alignment of the Slow *Sod* sequence of *D. melanogaster* and the Sod sequence of *D. simulans*. The *D. simulans* sequence is from KWIATOWSKI *et al.* (1989a). Translated parts of the exons are indicated with the dark bars, untranslated by light shading.