Molecular Evolution of Two Linked Genes, *Est-6* and *Sod*, in *Drosophila melanogaster*

Evgeniy S. Balakirev,*†,‡ Elena I. Balakirev,* Francisco Rodríguez-Trelles*‚§ and Francisco J. Ayala*

*Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, †Institute of Marine Biology, Vladivostok 690041, Russia, ‡Department of General Biology, Ecology and Soils, Far Eastern State University, Vladivostok 690600, Russia and §Departament de Genètica, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Manuscript received June 7, 1999
Accepted for publication July 22, 1999

ABSTRACT

We have obtained 15 sequences of *Est-6* from a natural population of *Drosophila melanogaster* to test whether linkage disequilibrium exists between *Est-6* and the closely linked *Sod*, and whether natural selection may be involved. An early experiment with allozymes had shown linkage disequilibrium between these two loci, while none was detected between other gene pairs. The *Sod* sequences for the same 15 haplotypes were obtained previously. The two genes exhibit similar levels of nucleotide polymorphism, but the patterns are different. In *Est-6*, there are nine amino acid replacement polymorphisms, one of which accounts for the *S-F* allozyme polymorphism. In *Sod*, there is only one replacement polymorphism, which corresponds to the *S-F* allozyme polymorphism. The transversion/transition ratio is more than five times larger in *Sod* than in *Est-6*. At the nucleotide level, the *S* and *F* alleles of *Est-6* make up two allele families that are quite different from each other, while there is relatively little variation within each of them. There are also two families of alleles in *Sod*, one consisting of a subset of *F* alleles, and the other consisting of another subset of *F* alleles, designed *F(A)*, plus all the *S* alleles. The *Sod* *F(A)* and *S* alleles are completely or nearly identical in nucleotide sequence, except for the replacement mutation that accounts for the allozyme difference. The two allele families have independent evolutionary histories in the two genes. There are traces of statistically significant linkage disequilibrium between the two genes that, we suggest, may have arisen as a consequence of selection favoring one particular sequence at each locus.

The understanding of the genome as an aggregate of relatively independent genes (bean-bag genetics) was a feature of the classical period of genetics. Gene interaction (epistasis) played a primary role in the theory of evolution starting in the 1920s (Wright 1931; Dobzhansky 1937; Schmalhausen 1946; Mayr 1953; Waddington 1957; Mather 1963). Linkage disequilibrium and nonrandom associations between alleles or groups of nucleotides may indicate epistatic relationships, and much empirical work has been devoted over several decades to demonstrate that linkage disequilibrium occurs between nonallelic genes. The issue of gene interaction is also important in connection with the long-lasting neutralist-selectionist controversy that has generated much research in population genetics for nearly 30 years (Kimura 1968, 1983;Kimura and Ohta 1971; Ayala et al. 1971, 1972a,b; Lewontin 1974; and many others). Linkage disequilibrium is often considered strong evidence supporting the selectionist position, especially if its pattern is consistent between populations (Lewontin 1974).

The evidence for linkage disequilibrium between individual loci remains scarce, except when genes are very closely linked or associated with chromosomal inversions (reviewed by Langley 1977; Hedrick et al. 1978; Barker 1979; Krimbas and Powell 1993). In the cases when significant associations have been detected, it is often far from clear whether they are caused by nonrandom haplotype sampling, random genetic drift, or natural selection (Mukai et al. 1974; Mukai and Voelker 1977). Significant disequilibrium can indeed arise without epistasis as a result of random genetic drift within a given population (Hill and Robertson 1968; Ohta and Kimura 1969a,b; Hill 1975, 1976), in subdivided populations (Nei and Li 1973; Li and Nei 1974; Feldman and Christiansen 1975; Ohta 1982a,b), and by founder effects (Avery and Hill 1979).

Numerous examples of significant linkage disequilibrium have been discovered in Drosophila between specific allozymes and chromosomal inversions, which have been interpreted as reflecting selection for favored multilocus allele combinations (Prakash and Lewontin 1968; Prakash 1974; Zouros 1976; Voelker et al. 1978). Ishii and Charlesworth (1977) and Nei and Li (1980) have, however, shown that nonrandom associations between allozymes and inversions can be ex-

Corresponding author: Francisco J. Ayala, Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525. E-mail: fjayala@uci.edu

Genetics 153: 1357-1369 (November 1999)
plained by absence or limited recombination without selection (as a consequence of chance and insufficient time for associations generated by mutation to decay by recombination and gene conversion). The inference from the allosyme studies is, therefore, that linkage disequilibrium is mostly associated with closely linked genes, but may involve distantly linked genes when special cytological mechanisms (polymorphic inversions) allow it to exist. Allozyme loci that can recombine freely exhibit little, if any, linkage disequilibrium. Failure to detect disequilibrium may, of course, be a consequence of the limited statistical power of the tests to detect it (Brown 1975), which might be overcome by using larger sample sizes and combining probabilities from independent tests (Brown 1975; Zapata and Alvar ez 1992, 1993).

The introduction of DNA sequencing and other molecular techniques in population studies makes it possible to gain considerable information about linkage disequilibrium. Nonrandom associations have been detected between polymorphic sites of Adh, Adh-Dup (Schaeffer and Miller 1993), and Est-5B (Veuille and King 1995) in Drosophila pseudoobscura; vermillion (Begin and Aquadro 1995) and G6pd (Eanes et al. 1996) in D. simulans; and the following in D. melanogaster: Adh and Adh-Dup (Kreitman and Hudson 1991), vermillion (Begin and Aquadro 1995), Pgd (Begin and Aquadro 1994), white (Kirby and Stephan 1995, 1996), G6pd (Eanes et al. 1996), dpp (Richter et al. 1997), Acp70A (Cirera and Aguadé 1997), period (Rosato et al. 1997), and Acp26Aa and Acp26Ab (Aguadé 1998). But, as in the case with allozymes, a positive correlation between the presence of chromosome inversions and the extent of linkage disequilibrium is often the case (Aquadro 1993, and references therein). Nevertheless, DNA linkage disequilibrium between nucleotide sites is well established in some cases, as well as the fact that it reflects epistatic relationships (Kirby et al. 1995; Kirby and Stephan 1996), although the nature of the epistatic interactions between genes remains enigmatic.

In this article, we investigate the nucleotide polymorphisms in the Est-6 and Sod genes of D. melanogaster and compare them in the two genes, seeking to identify processes that contribute to the polymorphisms. We test whether linkage disequilibrium may occur between these two fairly closely linked genes, as has been intimated by the results of SmiT -MBr i de et al. (1988), who investigated natural and laboratory populations and detected linkage disequilibrium between the allozyme polymorphisms of Sod and Est-6, but not between any other gene pairs. Hudson et al. (1994, 1997) have, moreover, shown that a selective sweep has recently occurred involving a many kilobases-long region that includes the Sod gene. Est-6 and Sod are closely linked on the left arm of chromosome 3 of D. melanogaster, 1000 kb apart (Procutier et al. 1991; Heino et al. 1994; Hartl and Lozovskaya 1995). The two genes are examined in a set of 15 haplotypes from a natural population in California.

**MATERIALS AND METHODS**

**Drosophila strains:** The 15 D. melanogaster strains were derived from wild flies collected by F. J. Ayala (October 1991) in El Rio Vineyard (Acampo, CA). The strains were made fully homozygous for the third chromosome by means of crosses with balancer stocks, as described by Seager and Ayala (1982). The homozygous strains were named in accordance with the Cu,Zn superoxide dismutase (SOD) electrophoretic allele they carry. Fast (F) or Slow (S), as follows: 255S, 438S, 510S, 521S, 94F, 174F, 377F, 483F, 498F, 521F, 565S, 581F, 968F, 517S, and 357F. The Sod gene sequence of these strains has been investigated previously in our laboratory (Hudson et al. 1994, 1997).

**Allozyme electrophoresis:** Twenty flies from each D. melanogaster strain were homogenized in 20 μl 0.1 M Tris-HCl buffer, pH 8.0. The homogenates were electrophoresed for 8–9 hr using a Trisborate-EDTA continuous buffer system, pH 8.6 (Macke r t and Faul hab er 1965) at 100 V in a 11% starch gel. The gels were stained for esterase-6 (EST-6) activity according to standard procedures (Man ch en ko 1994). See Ayala et al. (1972b) for additional details.

**DNA extraction, amplification, and sequencing:** Total genomic DNA was extracted using the procedure described by Palumbi et al. (1991).

We used the Est-6 sequence, previously published by Coll et al. (1990), for designing PCR and sequence primers. The amplified fragments (2017 bp long) included 56 bp of the 5′-flanking region, the Est-6 gene, the intergenic region, and 82 bp of the 3′-flanking region. The Est-6 gene has been referred to in the literature as Est-P, but the evidence indicates that it is a pseudogene (Balakirev and Ayala 1996).

The two primers used for the PCR amplification reactions (1 and 2; Figure 1) were 5′-gcaattgccagctcaagagtt-3′ (forward primer) and 5′-cagaatacacaagggctcatgct-3′ (reverse primer). All PCR reactions were carried out, as described by Kwiatowski et al. (1991), in final volumes of 100 μl containing 40 μM each of dNTP, 2.5 units of AmpliTaq DNA Polymerase (Perkin Elmer, Norwalk, CT), 0.2 μM each of forward and reverse primers, and 0.2 μM each of forward and reverse primers, buffer (Perkin Elmer) at a final concentration of 10 mm Tris-HCl (pH 8.3), 50 mm KCl, 1.5 mm MgCl₂, and ~50 ng of template (total genomic) DNA. The mixtures were overlaid with mineral oil, placed in a DNA thermal cycler (Perkin Elmer), incubated 5 min at 95°C, and subjected to multiple cycles of denaturation, annealing, and extension under the following conditions: 95°C for 1 min, 56°C for 1 min, and 72°C for 2 min (for the first cycle and progressively adding 3 sec at 72°C for every subsequent cycle). After 30 cycles, a final 7-min extension period at 72°C concluded each amplification reaction. Samples were stored at 4°C for several hours or at −20°C for up to 2 wk.

One-tenth of each reaction volume was assayed on a 0.8% agarose gel. If the desired PCR product was detected, the remainder volume of the reaction was purified with the Wizard PCR purification system (Promega, Madison, WI). The purified PCR product was directly sequenced by the dideoxy chain termination technique (Sanger et al. 1977) using Dye Terminator chemistry and separated with the ABI PRISM 377 automated DNA sequencer (Perkin Elmer). For each line, the sequences of both strands were determined. Eight internal primers spaced on average 350 nucleotides (a-d for the 5′-to-3′ direction and a′-d′ for the 3′-to-5′ direction; Figure 1) were used for sequencing the Est-6: (a) 5′-caccaaaattgtagga gagc-3′, (b) 5′-aactaggtcaagattgacctg-3′, (c) 5′-ctgtatggaagc-3′.
Cattanegat-3', (d) 5' gacttcacacataagtg-3', (a) 5' tcaagctacctcgctgtg-3', (b) 5' gctgctggagcatgaggtcctc-3', (c) 5' tcaagctccaaaaaggtcctc-3', and (d) 5' aaattccatacatcttcctactct-3'. The Est-6 sequences reported in this article have been deposited in the GenBank sequence database library under accession nos. AF147095-AF147102.

For the Sod gene, we analyzed a region (1408 bp long) that included 43 bp of exon I, the intron (725 bp), exon II (396 bp), and 244 bp of the 3' flanking region. The DNA preparation, amplification, cloning, and sequencing procedures for the Sod gene are described by Hudson et al. (1994, 1997).

DNA sequence analysis: All primers were designed using the computer program DNASIS for Windows (1994, Hitachi Software Engineering), which allows one to check the secondary structure of primers. Multiple alignment was carried out manually, using the program DARWIN (elaborated by Robert Tyler from our laboratory), and automatically, using the program CLUSTAL W (Thompson et al. 1994). Two DNA sequences were obtained from the GenBank database: D. melanogaster, Est-6 (accession no. M33780), and D. simulans, Sod (accession no. X15685). The maximum parsimony analysis of the Est-6 and Sod sequences was performed with the computer program PAUP (Swoford 1993). The computer program DNA SP (Rozas and Rozas 1997) was used to analyze the data by means of a "sliding window" (Hudson and Kaplan 1988) and was used for most intraspecific analyses.

Linkage disequilibrium analysis: Linkage disequilibrium within and between Est-6 and Sod was evaluated using Fisher's exact test and chi-square test for independence between sites. Singleton polymorphisms (mutations appearing in only one sampled allele) were omitted. Lewontin's sign test, elaborated especially for molecular sequence data (Lewontin 1995), was also used to test the significance of linkage disequilibrium between Est-6 and Sod, including singletons. Lewontin (1995) has shown that Fisher's exact test is not sensitive in cases with very asymmetrical distribution of allele frequencies, a situation that is commonly observed in molecular sequence data, including our data. In such a situation, there is, for numerical reasons, very low probability of obtaining a significant test by means of the usual two-by-two contingency tables (Fisher's exact test or chi-square test). To overcome this problem, Lewontin's (1995) test is based on the distribution of the disequilibrium sign ("sign" test), which allows the analysis of asymmetrical allele frequency data to make inferences about overall linkage disequilibrium. The procedure involves examining the number of positive and negative D values for each polymorphic site within and between all types of pairwise comparisons (singletons vs. singletons, singletons vs. doublets, doublets vs. doublets, and so on; Lewontin 1995). The observed value, summed for a particular type of pairwise comparison, is compared with the expected value using a goodness-of-fit test (the likelihood ratio statistic, G-test).

RESULTS

The genes Est-6 and Sod are on the left arm of chromosome 3 of D. melanogaster, genetically mapped at 35.9 and 32.5 and located at 69A1-A5 and 68A8-A9 on the polyteny chromosomes, respectively (Heino et al. 1994).

Allozyme polymorphism: We have analyzed Est-6 in 15 D. melanogaster strains, fully homozygous for the third chromosome, derived from flies collected in the El Rio Vineyard in California. The strains were chosen because their Sod gene sequences have been previously investigated in our laboratory (Hudson et al. 1994, 1997) and may be considered as randomly selected with respect to their Est-6 alleles. Electrophoretic analysis reveals two common EST-6 allozymes: Fast (357F and 517S) and Slow (2555, 4385, 5105, 5215, 94F, 174F, 377F, 483F, 498F, 521F, 565F, 581F, and 968F strains). (The S or F in each strain's designation refers to the SOD allele it carries; see materials and methods.)

Nucleotide polymorphism: The organization of the Est-6 gene is outlined in Figure 1. The sequenced region is 1879 bp long, comprising the Est-6 gene and the 193 bp of the intergenic region between Est-6 and Sod.

Table 1 displays the gene lengths, the number of polymorphic sites, and the number of polymorphic sites, 9 of which yield amino acid replacements.

Table 2 gives the values of η and θ for the two genes, Est-6 and Sod. The data for Sod are obtained from Hudson et al. (1997). The overall nucleotide diversity is somewhat larger for Sod. However, the diversity is about the same for the coding or the noncoding regions considered separately (Sod has greater overall diversity because the noncoding region is larger). The diversity, as expected, is much greater in the noncoding than in the coding regions, reflecting the usual higher level of functional constraint in the coding regions of genes. The distribution of polymorphic sites among the Sod haplotypes is extremely nonuniform: 69.6% of the Sod polymorphic sites are introduced by just two haplotypes (498F and 968F; see Hudson et al. 1994, 1997). It is also quite polarized in the case of Est-6, with 62.9% of the polymorphic sites resulting from two haplotypes, 517S and 357F, which are the only two Est-6 F alleles in the sample (shown at the bottom of Table 1). When these two pairs of sequences are removed, the η and θ values considerably decrease for both genes, but in the coding region, the decrease is somewhat greater for Est-6 (Table 2). This difference may be due to different selective constraints in the two genes, but it may also reflect the different histories of the particular sets of alleles included in our analysis (see discussion).

There are 26 polymorphic nucleotide sites (1.6%) in the coding region of Est-6, but only 5 (1.1%) in Sod. In Est-6, the transversion/transversion (Tv/Tv) ratio is 3/23 = 0.130, much lower than expected from random mutation, reflecting the usual selection effect againstTv. The Tv/Ts ratio for Sod is significantly higher than for Est-6 (2/3 = 0.667). The ratio of replacement to synonymous segregating sites is lower for Sod (1/4 = 0.250) than for Est-6 (9/17 = 0.529). The Sod data of our study do not come from a random sample, since the number of S alleles was made intentionally larger (33%) in the sample than would be expected in a random sample (frequency of S in the natural population is ~5–15%; see Hudson et al. 1994, 1997). Thus, in our sample, there are more replacement substitutions in Sod than would be expected and fewer synonymous substitutions, since the S alleles are all completely (or very
D. melanogaster tends to have the highest incidence of Moriya and Powell (1996) indicate that D. melanogaster tends to have the highest incidence of replacement polymorphisms (26.4%) when compared

Haplotype structure: Figure 2 represents the maximum parsimony tree of the Est-6 haplotypes. The distribution of the pairwise differences is bimodal, owing to haplotypes 517S and 357F, which are largely different from the rest, although quite similar to each other. These two haplotypes code for the Est-6 Fast allozyme and will be denoted as the Est-6 F allelic lineage, whereas the other 13 haplotypes will be denoted as the Est-6 S allelic lineage. The great divergence between the two lineages indicates that this S-F enzyme polymorphism is ancient, at least relative to the allelic diversity within each of the two lineages.

In D. melanogaster, the Est-6 S → F (asparagine → aspartic acid) replacement is associated with site 772 (A → G). D. simulans has an A at position 772, which suggests that in D. melanogaster, the S allele may have been the ancient condition from which the F allelic lineage derived. However, Cooke and Oakeshott (1989) and Hassan and Eanes (1996) suggest that the F lineage is ancestral on the grounds that the overall level of polymorphism is significantly greater among the other two Drosophila species, D. simulans (11.6%) and D. pseudoobscura (16.9%).

For Est-6, the A → G replacement substitution (nucleotide site 772, Table 1) results in a charge-altering amino acid replacement (Asn → Asp, amino acid position 258), which was first detected by Cooke and Oakeshott (1989). Cooke and Oakeshott (1989) have suggested that another replacement substitution (G → A, at nucleotide site 802, resulting in Ala → Thr at amino acid site 268) might also contribute to the selective differences observed between the two Est-6 allozymes.

The reference sequence is from Collett et al. (1990). The 15 D. melanogaster haplotypes are designed according to the Sod strain from which they originate. The numbers above the sequences represent the positions of segregating sites. Nucleotides are numbered from the initiation codon. The coding regions (exon I and exon II) are underlined below the reference sequence. Amino acid replacement polymorphisms are marked with asterisks. The SF allozyme polymorphism is determined by site 772, where S has A (asparagine) and F has G (aspartic acid). The nucleotides in the D. simulans sequence are shown at the bottom; only those sites that are polymorphic in D. melanogaster are shown.

The reference sequence is from Collett et al. (1990). The 15 D. melanogaster haplotypes are designed according to the Sod strain from which they originate. The numbers above the sequences represent the positions of segregating sites. Nucleotides are numbered from the initiation codon. The coding regions (exon I and exon II) are underlined below the reference sequence. Amino acid replacement polymorphisms are marked with asterisks. The SF allozyme polymorphism is determined by site 772, where S has A (asparagine) and F has G (aspartic acid). The nucleotides in the D. simulans sequence are shown at the bottom; only those sites that are polymorphic in D. melanogaster are shown.
**TABLE 2**

**Nucleotide diversity of Sod and Est-6 in 15 strains of D. melanogaster**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full sequence</th>
<th>Coding regions</th>
<th>Noncoding regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \pi )</td>
<td>( \theta )</td>
<td>( \pi )</td>
</tr>
<tr>
<td>All 15 haplotypes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est-6</td>
<td>1879 (1.63)</td>
<td>5.73</td>
<td>1635 (1.38)</td>
</tr>
<tr>
<td></td>
<td>4.50 (0.97)</td>
<td></td>
<td>5.73 (0.96)</td>
</tr>
<tr>
<td>Sod</td>
<td>1408 (2.58)</td>
<td>9.45</td>
<td>439 (0.76)</td>
</tr>
<tr>
<td></td>
<td>6.67 (1.44)</td>
<td></td>
<td>9.45 (1.57)</td>
</tr>
<tr>
<td>Excluding the 2 F haplotypes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est-6</td>
<td>1879 (0.46)</td>
<td>2.23</td>
<td>1635 (0.38)</td>
</tr>
<tr>
<td></td>
<td>1.51 (0.62)</td>
<td></td>
<td>2.23 (0.62)</td>
</tr>
<tr>
<td>Sod</td>
<td>1408 (0.69)</td>
<td>3.43</td>
<td>439 (0.45)</td>
</tr>
<tr>
<td></td>
<td>2.17 (0.89)</td>
<td></td>
<td>3.43 (1.27)</td>
</tr>
</tbody>
</table>

\( \pi \) is the average number of nucleotide differences per site among all pairs of sequences (Nei 1987, p. 256). \( \theta \) is the average number of segregating nucleotide sites among all sequences, based on the expected distribution of neutral variants in a panmictic population at equilibrium (Watterson 1975). Standard deviations (SD) are given in parentheses below the parameter values. All \( \pi \), \( \theta \), and SD values are multiplied by 10\(^3\). Values for all 15 haplotypes are on top. Below are the values obtained after excluding the two F alleles for Est-6 (strains 357F and 517S) and the 498F and 968F strains for Sod.

F than among S alleles. The two Est-6 allelic lineages (F and S) are about equally different from D. simulans (90–98 nucleotide substitutions per sequence), much more so than they are from each other (22–29 substitutions per sequence), which indicates that the F-S polymorphism originated well after the divergence of the two species. The haplotypes have accumulated a number of substitutions after the divergence between the F and S allelic lineages: the two F strains differ by 4 synonymous substitutions, but no replacements, whereas there are 4 polymorphic replacement sites among the 13 S haplotypes, 1 of which (at 1526) is shared by 2 haplotypes (94F and 174F), for a total of 5 replacement polymorphisms among the 13 S haplotypes.

Figure 3 represents the maximum parsimony tree for the Sod haplotypes. The contrasts with Est-6 are notable. There is only one replacement polymorphism (F → S allozyme) in Sod. Two F haplotypes (498F and 968F) are very different from all others, whereas the eight other F haplotypes are much more similar to the S haplotypes than they are to 498F and 968F.

**Linkage disequilibrium:** Within Est-6, 262 out of 351 pairwise comparisons (74.6%) between nonsingleton pairs of polymorphisms show statistically significant linkage disequilibrium by the chi-square test; with the Bonferroni correction for multiple comparisons, there are 192 (54.7%) significant associations. The distribution of significant associations is fairly uniform across the Est-6 sequence; linkage disequilibrium does not decline as distance between polymorphic sites increases.

We have also found an excess of nonrandom associations within Sod: 211 out of 325 pairwise comparisons (64.9%) are significant, and 191 (58.8%) are significant with the Bonferroni correction. The significant associations do not form any obvious cluster, nor is the strength of linkage disequilibrium related to the distance between polymorphic sites.

We have first evaluated linkage disequilibrium between the Sod and Est-6 genes using Fisher’s exact test and the chi-square test, which fail to detect any significant interlocus association, as might be expected owing to asymmetrical allelic frequencies (Lewontin 1995;
It is interesting that in this last comparison, the (Figure 4B). In both cases, there is a distinct peak of

$G_{Est-6}$ and the $3_{Est-6}$ have applied the Lewontin test separately to different using the Tajima test ($D_G$) i.e., $Rohlf$ of negative associations is 1551

The observed pattern of linkage disequilibrium between the two genes remains unchanged when singletons are excluded.

**Test of neutrality:** We have applied the HKA (Hudson et al. 1987), Tajima (1989), McDonald and Kreitman (1991), and Fu and Li (1993) tests of neutral equilibrium to our sequence data. We have first examined the intergenic region between Est-6 and $q_{Est-6}$ vs. the coding sequence of Est-6 without observing any significant departures from neutral expectations within D. melanogaster relative to the differences between D. melanogaster and D. simulans ($\chi^2 = 1.24$, $P > 0.05$). The same result was obtained by Karotam et al. (1995), who used various upstream sequences of Est-6 and Adh as neutral reference sequences and various pairs of D. melanogaster and D. simulans Est-6 alleles. Moriyama and Powell (1996) found significant deviation from neutral expectations in the comparison of the intraspecific polymorphism of Est-6 with the interspecific divergence between Est-6 and Pgd. However, nucleotide variation at Pgd is highly unusual because most HKA tests that include this locus (whether as the reference or test locus) show significant deviations from neutrality (Moriyama and Powell 1996). We have applied the Tajima (1989) and Fu and Li (1993) tests separately to different parts of Est-6 and to the whole gene. These tests do not reveal any significant deviation from neutrality. However, the Tajima (1989) test applied to our data (exon I), combined with previously published Est-6 sequences (Cooke and Oakeshott 1989; Hasson and Eanes 1996), reveals significant deviation from neutrality expectations for the S alleles ($D = -1.864$, $P < 0.05$), but not for the F alleles ($D = -0.181$, $P > 0.10$). Also, the McDonald and Kreitman (1991) test applied to four D. simulans (from Karotam et al. 1995) and 15 D. melanogaster Est-6 coding sequences reveals significant deviation from neutrality ($G = 7.07$, $d.f. = 1$, $P < 0.01$, Table 4). The ratio of replacement to silent polymorphism within species is lower than the ratio of fixed replacement to silent differences between them.

We have applied all the neutrality tests mentioned above to the Sod data from Hudson et al. (1997) and, using the Tajima test ($D = -1.873$, $P < 0.05$), found significant departure from neutrality only for the $3'$-flanking region.

**Sliding window analysis:** We have analyzed separately different regions of Est-6 and Sod with the sliding window method (Hudson and Kaplan 1988). Figure 4 shows the sliding window plots of exon I of Est-6 for all sequences (Figure 4A) and for the S allelic lineage only (Figure 4B). In both cases, there is a distinct peak of increased variation in the region surrounding the replacement site (772) responsible for the Est-6 F-S allo-


**TABLE 3**

Linkage associations for pairs of biallelic sites at the *Est-6* gene

<table>
<thead>
<tr>
<th>(k)</th>
<th>(m)</th>
<th>(D^{+}) Observed</th>
<th>(D^{+}) Expected</th>
<th>(D^{-}) Observed</th>
<th>(D^{-}) Expected</th>
<th>(G)</th>
<th>(G^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>8.576</td>
<td>122</td>
<td>119.424</td>
<td>0.933</td>
<td>0.929</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6</td>
<td>74.480</td>
<td>554</td>
<td>485.520</td>
<td>115.968</td>
<td>115.865</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5</td>
<td>11.200</td>
<td>45</td>
<td>35.184</td>
<td>13.434</td>
<td>13.296</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>12.816</td>
<td>45</td>
<td>35.184</td>
<td>13.434</td>
<td>13.296</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2.664</td>
<td>4</td>
<td>5.336</td>
<td>0.947</td>
<td>0.891</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>147.004</td>
<td>570</td>
<td>424.996</td>
<td>317.465</td>
<td>317.188</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>43.778</td>
<td>116</td>
<td>74.222</td>
<td>91.251</td>
<td>90.866</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3.096</td>
<td>3</td>
<td>2.904</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5.733</td>
<td>6</td>
<td>3.267</td>
<td>3.409</td>
<td>3.230</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1.472</td>
<td>0</td>
<td>0.528</td>
<td>1.226</td>
<td>0.981</td>
</tr>
</tbody>
</table>

\(k\) is the number of copies of the rarer allele at a biallelic site; \(m\) is the number of copies of the rarer allele at another biallelic site for \(m = k\). \(D^{+}\) and \(D^{-}\) refer to positive and negative associations, respectively.

zyme polymorphism; the peak becomes more apparent when only the S allelic lineage is considered (Figure 4B). The same peak is also clearly distinguishable for the data of Cooke and Oakeshott (1989) and Hasson and Eanes (1996) (Figure 5, A and B). For the Sod gene, a distinct peak appears within the intron (Figure 6A) and toward the end of the 3’-flanking region (Figure 6B). These two peaks coincide with areas of significantly high values, according to Tajima’s D-test.

Overall, the sliding window analysis and the neutrality tests (Tajima’s and McDonald and Kreitman’s) suggest that the polymorphism distribution in the *Est-6* and *Sod* genes significantly deviates from the expectations of neutrality.

**Interspecific comparisons and divergence time:** The *Est-6* average distance (nucleotide differences) between *D. simulans* and *D. melanogaster* is 91.9 (see Table 4). The average distance between the two main (F and S) allelic lineages of *D. melanogaster* is 25.5. If we assume that the divergence between the species occurred 2.3–2.5 mya (Powell and DeSalle 1995; Russo et al. 1995), the divergence of the two *D. melanogaster* lineages would have occurred ~666,000 years ago—assuming that *Est-6* is a good molecular clock. The divergence between the two F alleles occurred ~105,000 years ago and between S alleles ~73,000 years ago.

The Sod average distance between *D. melanogaster* and *D. simulans* is 73.1, while between the two main allelic lineages [F vs. F(A)S] it is 29.3, which corresponds to ~962,000 years, while the divergence between the two F alleles occurred ~427,000 years ago. The average distance among all F(A) and S alleles (excluding 581F, which is probably a recombinant) is 2.4, corresponding to 79,000 years ago; the average distance among the S alleles is 0.8 or ~26,000 years ago.

We have calculated the time of divergence between and within the allelic lineages for both genes following the Hudson et al. (1997) approach, using all sequences or their homogeneous subsets [see Hudson et al. (1997) for details]. For the *Est-6*, the expected time of divergence between the F and S allelic lineages is ~223,000 years \[\mu \times t \times 15 \times 15 = 35; t = 223,000 \text{ years}\], where \(\mu\) is the neutral mutation rate at non-coding and silent sites, assumed to be \(16 \times 10^{-9}\) per

**TABLE 4**

McDonald-Kreitman test of neutrality for the *Est-6* gene

<table>
<thead>
<tr>
<th>Polymorphisms</th>
<th>Fixed differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D. <em>melanogaster</em></td>
</tr>
<tr>
<td>Replacement</td>
<td>19</td>
</tr>
<tr>
<td>Silent</td>
<td>25</td>
</tr>
</tbody>
</table>

\(^a\) Calculated from Karotam et al. (1995).

\(^b\) Sites that are polymorphic in both species are counted only once. For the two-tailed Fisher's exact test, \(P = 0.013; G\)-test without correction = 7.070, \(P = 0.008; G\)-test with Williams' correction = 6.949, \(P = 0.008; G\)-test with Yates' correction = 6.042, \(P = 0.014.\)
site per year. For the Sod F(A)S and F allelic lineages, the divergence time is $\sim 178,000$ years [$\mu \times t \times 15 (1408 - 0.75 \times 439) = 46; t = 177633.6$]. The divergence time within the Est-6 S allelic lineage is $96,000$ years [$\mu \times t \times 13 (1879 - 0.75 \times 1635) = 13; t = 95748.8$], considering all polymorphic sites, and $59,000$ years [$\mu \times t \times 13 (1879 - 0.75 \times 1635) = 8; t = 58922.3$], if we exclude putative recombinant sites. The expected time of divergence within the Sod F(A)S allelic lineage is $\sim 45,000$ years [$\mu \times t \times 13 (1408 - 0.75 \times 439) = 10; t = 44556.9$] and $27,000$ years [$\mu \times t \times 13 (1408 - 0.75 \times 439) = 6; t = 26734.1$], excluding the putative recombinant sites.

**DISCUSSION**

The Est-6 and Sod genes of D. melanogaster are closely linked on the left arm of chromosome 3, separated by $3.4$ cM, or $1$ Mb. Both genes are characterized in natural populations by a polymorphism with two common alleles, S and F, which differ by a single-nucleotide substitution and a corresponding amino acid replacement. In Sod, the S-F replacement is the only amino acid polymorphism commonly found in natural populations, whereas in Est-6, additional rare amino acid replacements are found.

**Natural selection:** The Cu,Zn SOD is involved in protecting the cell against the toxicity of oxygen radicals by scavenging superoxide radicals and dismutating them to hydrogen peroxide and molecular oxygen (Fridovich 1986). The Sod F-S polymorphism is of recent evolutionary origin, as shown by the virtually complete absence of silent variation among S alleles collected in populations that are geographically very distant, from China and Europe to the United States (Hudson et al. 1994, 1997). S alleles have not been found in South America, nor in Africa, whence D. melanogaster spread throughout the world, probably in recent millennia. In the United States, the frequency of S has been found to range from 0 to 0.15 in different populations, and can vary from year to year in a given population from 0.05 to 0.15 (Hudson et al. 1994, 1997). The S alleles are completely or very nearly identical in nucleotide sequence to a set of F alleles denominated F(A). The F(A) alleles have not been found in Africa, but are present in Europe, the United States, and South America, where they often account for $\sim 50\%$ of all F alleles in a population (Hudson et al. 1994, 1997).

![Figure 4](image1.png)

**Figure 4.**—Sliding window plot of exon I polymorphism ($\pi$) in the Est-6 gene of D. melanogaster. (A) All data. (B) Excluding the F allelic lineage. Window size is 100 nucleotides, with one-nucleotide increments.

![Figure 5](image2.png)

**Figure 5.**—Sliding window plot of exon I polymorphism ($\pi$) in the Est-6 gene of D. melanogaster. (A) Data of Cooke and Oakeshott (1989). (B) Data of Hasson and Eanes (1996). Window size is 100 nucleotides, with one-nucleotide increments.
A hypothesis of the geographic evolution of the Sod alleles consistent with the information just summarized is as follows. The F(A) mutation arose in a D. melanogaster population outside Africa ~5000 years ago (Hudson et al. 1997) and rapidly spread throughout other continents, impelled by natural selection (Hudson et al. 1994). The rapidity of the F(A) world expansion is evidenced by the virtually complete absence of silent substitutions throughout a fragment >10 kb that includes Sod (Hudson et al. 1997). The S allele arose by a single-nucleotide replacement from F(A), either in Europe or the United States, and spread from one to the other continent, but never reached South America or Africa. The nucleotide identity of S alleles from widely distant continents (and the nucleotide identity of a fragment >10 kb that includes Sod S) favors the interpretation that the F(A) → S mutation occurred only once. The rapid expansion of S in Europe and the United States must have been impelled by natural selection. It is uncertain, however, whether the selective advantage is the same favoring F(A) over other F alleles, or whether the S replacement is also favored. The S and F enzymes differ in such biochemical properties as thermostability and specific activity (Lee et al. 1981). Moreover, the S allele is at an advantage relative to F in heavily irradiated populations (Peng et al. 1991) or in those selected for reproduction at an advanced age (Tyler et al. 1993).

Detecting selection is handicapped in our sample because we have studied only 15 sequences (Hansing et al. 1994; Simonsen et al. 1995; Hasson and Eanes 1996; Richter et al. 1997). Nevertheless, we have found evidence of natural selection within the 3'-flanking region of Sod with Tajima's (1989) test. Moreover, the sliding window method of Hudson and Kaplan (1988) manifests a distinct peak in the 3'-flanking region as well as within the intron of Sod; these peaks are expected as a consequence of balancing selection (Strobeck 1983; Hudson and Kaplan 1988). Selective effects within introns have been recognized previously. Berry and Kreitman (1993) have shown that an intron polymorphism at Adh in D. melanogaster exhibits a more pronounced cline than the F-S alolzyme polymorphism that is generally attributed to selection at that locus. Kirby and Stephan (1995) have proposed that positive selection accounts for intron polymorphism also at the white locus of D. melanogaster. This may be because introns may include regulatory sequences (Bingham et al. 1988; Aronow et al. 1989; Gasch et al. 1989; Huang et al. 1993; Pogulis and Freytag 1993). There is also other evidence for selective constraints and epistatic selection on the nucleotide sequence evolution of introns (Learn et al. 1992; Leicht et al. 1993, 1995; Stephan and Kirby 1993; Kirby et al. 1995).

The Est-6 protein is transferred by males to females in the semen fluid during copulation in D. melanogaster (Richmond et al. 1980; Richmond and Senior 1981), and it affects the female's consequent behavior and mating proclivity (Gromko et al. 1984; Scott 1986). The Est-6 coding sequence is selectively constrained (Cooke and Oakeshott 1989; Karotam et al. 1993, 1995), although presumably to a lesser extent than Sod, which exhibits only one polymorphic amino acid site in our sample vs. the nine polymorphic sites present in Est-6. The Est-6 5'-flanking region contains positive cis-regulatory elements controlling the expression of Est-6 and may contain binding sites for regulatory protein factors involved in tissue-specific transcription control. The evidence indicates that the 5'-flanking region is evolving under greater selective constraints than the coding region (Karotam et al. 1993, 1995; Odgers et al. 1995). The occurrence of parallel latitudinal clines of the Est-6 F-S polymorphism in Europe, North America, and Australia, the pattern of temporal variation, and other lines of evidence all indicate that the Est-6 F-S polymorphism is subject to balancing selection (Oakeshott et al. 1989, 1993, 1995; Richmond et al. 1990). The peak around the S-F amino acid polymorphism we have observed (Figure 4) is also evidence of balancing selection impacting the Est-6 S-F polymorphism in our sample. The detected tendency cannot be explained by chance, because similar peaks are observed in the Est-6 data of other authors (Cooke and Oakeshott et al. 1989, 1993, 1995).
that the selective pressure favoring Est-6. This scenario is consistent with the interpretation allowing for only limited recombination between Sod and Est-6. The fact that the two Sod F alleles that are not F(A) occur in haplotypes with Est-6 S alleles is not surprising because the Est-6 S allozyme is the most ancient and common.

An alternative historical scenario that might account for the presence of the two largely divergent sets of alleles that we observe at each locus is population subdivision with subsequent admixture. D. melanogaster may have been geographically split into two populations that remained separate for a time long enough to accumulate a number of nucleotide substitutions within each gene. The substitutions would have accumulated independently in the two populations. Recent admixture of the two populations would have brought together the two sets of alleles, as we find them in the El Rio population, where our strains were collected. According to this scenario, however, the two sets of alleles at the two loci would be associated in the same haplotypes. This is not the case. As shown in Figures 2 and 3, the strains with the two Est-6 F alleles (517S and 357F) are different from the strains carrying the two Sod F alleles (377F and 581F).

**History of the allelic lineages:** Figures 2 and 3 are maximum-parsimony trees of the Est-6 and Sod alleles. It is apparent that the phylogeny of the electrophoretic alleles is different in the two genes. The two Est-6 F alleles are quite similar to each other, but they are in haplotypes that carry an F(A) Sod allele in one case and an S Sod allele in the other case. The 13 Est-6 S alleles are associated with S or F Sod alleles, and the Sod F alleles include closely related F(A) alleles as well as the other distant F alleles. The F(A) and S alleles of Sod have diverged very recently (Hudson et al. 1994, 1997), which is also apparent in Figure 3 and our extensive unpublished data (we reckon that the exceptional Sod allele 581F represents a case of intragenic recombination). We did not, however, expect that the two Est-6 F alleles would be closely related to one another, and even less so that the 13 Est-6 S alleles would have quite similar (and even identical) nucleotide sequences. Rather, we would have expected that the Est-6 S alleles, which may represent the ancestral electrophoretic state, would consist of alleles quite heterogeneous in nucleotide sequence. Such is, indeed, the case for the Sod alleles, which, if we exclude the F(A) set, are extremely heterogeneous in nucleotide sequence (Hudson et al. 1994; F. J. Ayala, unpublished data). A reasonable account of the observations is to assume that the Sod F(A) mutation occurred in a haplotype carrying an Est-6 S allele, and that the rapid world expansion of the Sod F(A) and S alleles greatly enhanced the frequency of that particular Est-6 S allele. The Sod F(A) mutation is estimated to have occurred ~5000 years ago or somewhat earlier (Hudson et al. 1997). The 5000 years of the world expansion of the Sod F(A) and S alleles would allow for only limited recombination between Sod and Est-6. This scenario is consistent with the interpretation that the selective pressure favoring Sod F(A) and S alleles has been strong. The two strains 517S and 357F [which have F(A) and S alleles, respectively, at Sod, but the Est-6 F allele] would have arisen by a recombination event between Sod and Est-6. The fact that the two Sod F alleles that are not F(A) occur in haplotypes with Est-6 S alleles is not surprising because the Est-6 S allozyme is the most ancient and common.

**Linkage disequilibrium:** Sod and Est-6 are 3.4 cM apart in the recombination map (3–32.5 and 3–35.9; Heino et al. 1994), separated by 983 kb in the genomic map (A. Long, personal communication). Yet the nucleotide polymorphisms are negatively associated between the two loci (Table 3, P < 0.001); that is, minority nucleotide substitutions (present in one to five strains) at one locus are negatively associated with minority nucleotide substitutions at the other locus. A selection account of this observation implies that mutations occurring at one locus are selected against if nucleotide substitutions are present in the other locus, but not if the other locus exhibits the consensus sequence.

There is an extensive genetic literature advancing the notion that the unit of evolution is not the single gene, but rather, interacting gene complexes (Wright 1931; Dobzhansky 1937; Schmalhausen 1946; Mathers 1953; Waddington 1957; Mayr 1963; Franklin and Lewontin 1970). There also are traditional arguments propounding that sets of genes for quantitative traits are built up in an alternated arrangement of plus and minus alleles on chromosomes, so that selection minimizes actual variation for such traits, but maximizes potential variability, which is released by recombination. This +/− alternating arrangement explains why populations rapidly respond to varying environmental challenges, as well as the success of artificial selection favoring extreme values of the traits (Mathers 1953; Thoday et al. 1964). However, even if we accept such theoretical constructs and their evidential support, it remains obscure why the plus/minus compensatory arrangement would occur between two genes that are
unlikely to share distinctive functional interactions, and between nucleotide substitutions that are in most cases synonymous.

It seems more likely that linkage disequilibrium has arisen as a consequence of selection strongly favoring one particular sequence at a locus, such as seems to have occurred at Sod, where the F(A) alleles (and the derivative S alleles) rapidly increased in frequency. Rare substitutions present in Est-6 would have hitchhiked along with Sod F(A) without allowing enough time for their elimination by purifying selection with or without recombination between the two loci. The reciprocal situation would have also occurred, in which common Est-6 alleles are favored by selection and low-frequency Sod substitutions are hitchhiking along. As noted earlier, there is evidence of positive selection in favor of Sod F(A); in the case of Est-6, the evidence favors balancing selection between the two common alleles, S and F. In a study of two wild samples and four experimental populations, Smit-McBride et al. (1988) did not observe linkage disequilibrium between Est-6 and Sod in the large wild samples, but linkage disequilibrium appeared after 8–30 generations in three of the four experimental populations derived from the wild samples. Population bottlenecks and other factors were excluded. Smit-McBride et al. (1988) noted that natural selection acting on the allozymes, or on loci very tightly linked to them, was the most likely explanation for the disequilibrium.

Hudson et al. (1997) have estimated that the time since the selective sweep for Sod F(A)S is ~5000 years or more, but not likely more than 30,000–40,000 years. Assuming a molecular (neutral) clock, the age of this lineage is estimated at 79,000 years. Similarly, the time of divergence between the two main Sod allelic lineages is 178,000 years with the method of Hudson et al. (1997), but five times greater, 962,000 years, under the molecular clock assumption. These discrepancies can be accounted for by natural selection accelerating the evolution of the selectively favored F(A)S alleles. Hudson et al. (1997) argue that the selection is strong (s > 0.01). Parallel differences occur for Est-6, but the differences between the estimates obtained by the two methods are smaller. Thus, the age of the Est-6 S allelic lineage is 60,000–95,000 vs. 73,000 years, and the divergence between the S-F lineages is 220,000 vs. 666,000 years, corresponding to the method of Hudson et al. (1997) vs. the clock method. This is consistent with the hypothesis that the divergence between the two Est-6 allelic lineages has been impelled by natural selection, but with lesser strength than in the case of Sod.

We thank several members of our laboratory: Kevin Bailey, Heather Carstens, Victor DeFilippis, Alberto García Sáez, Jan Kwiatowski, Carlos Machado, Stephen Rich, Douglas Skaecky, and Andrei Tatarinov for encouragement and help. Walter M. Fitch, Brandon Gaut, Richard R. Hudson, and Anthony Long read the manuscript and offered valuable comments. This work is supported by National Institutes of Health grant GM-42397 to F.J.A.

LITERATURE CITED


Prakash, S., 1974 Gene differences between the sex ratio and standard gene arrangements of the X chromosome and linkage disequilibrium between loci in the standard gene arrangement of the X chromosome in Drosophila pseudoobscura. Genetics 77: 795−804.
Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics 103: 545−555.
Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97−159.