Title
Nucleotide Variation of the Est-6 Gene Region in Natural Populations of Drosophila melanogaster

Permalink
https://escholarship.org/uc/item/3bk8p21s

Journal
Genetics, 165(4)

ISSN
0016-6731

Authors
Balakirev, ES
Ayala, FJ

Publication Date
2003-12-01

License
CC BY 4.0

Peer reviewed
Nucleotide Variation of the *Est-6* Gene Region in Natural Populations of *Drosophila melanogaster*

Evgeniy S. Balakirev*†‡ and Francisco J. Ayala*†

*Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, †Institute of Marine Biology, Vladivostok 690041, Russia and ‡Academy of Ecology, Marine Biology and Biotechnology, Far Eastern State University, Vladivostok 690600, Russia*

Manuscript received February 27, 2003
Accepted for publication August 20, 2003

**ABSTRACT**

We have investigated nucleotide polymorphism in the *Est-6* gene region in four samples of *Drosophila melanogaster* derived from natural populations of East Africa (Zimbabwe), Europe (Spain), North America (California), and South America (Venezuela). There are two divergent sequence types in the North and South American samples, which are not perfectly (North America) or not at all (South America) associated with the *Est-6* allele variation. Less pronounced or no sequence dimorphism occurs in the European and African samples, respectively. The level of nucleotide diversity is highest in the African sample, lower (and similar to each other) in the samples from Europe and North America, and lowest in the sample from South America. The extent of linkage disequilibrium is low in Africa (1.23% significant associations), but much higher in non-African populations (22.59, 21.45, and 37.68% in Europe, North America, and South America, respectively). Tests of neutrality with recombination are significant in non-African samples but not significant in the African sample. We propose that demographic history (bottleneck and admixture of genetically different populations) is the major factor shaping the nucleotide patterns in the *Est-6* gene region. However, positive selection modifies the pattern: balanced selection creates elevated levels of nucleotide variation around functionally important (target) polymorphic sites (*Rso*−/*Rso*+ in the promoter region and F/S in the coding region) in both African and non-African samples; and directional selection, acting during the geographic expansion phase of *D. melanogaster*, creates an excess of very similar sequences (*Rso*− and S allelic lineages, in the promoter and coding regions, respectively) in the non-African samples.

FROM the very beginning of the “allozyme era,” esterase 6 (*Est-6*) has been one of the most investigated and informative molecular markers in *Drosophila* population, evolutionary, and development genetics (reviewed by Oakeshott *et al.* 1989, 1993, 1995; Korochkin *et al.* 1990; Richmond *et al.* 1990). Wright (1961, 1963) described two main alleles (*Fast* and *Slow*) of *EST-6*, showed their Mendelian inheritance, found a differential response to the organophosphate inhibitor, and raised questions about the adaptive significance of the polymorphism. The main alleles show large-scale latitudinal clines (Oakeshott *et al.* 1981), with the Slow allele more common at higher latitudes. This, together with other data on the temporal and geographic allele variation in natural populations and results of laboratory experiments, suggests that the *EST-6* polymorphism is maintained by some form of positive selection (reviewed by Oakeshott *et al.* 1989, 1993, 1995; Richmond *et al.* 1990).

The *Est-6* gene is on the left arm of chromosome 3 of *Drosophila melanogaster*, at cytogenetic map position 69A1–A3 (Procunier *et al.* 1991). Oakeshott *et al.* (1987) first obtained the nucleotide sequence and characterized the exon-intron structure of the *Est-6* gene. Using available information on nine other eukaryotic esterases, they identified the active site and other functionally important regions of the gene. The coding region of *Est-6* is 1686 bp long and consists of two exons (1387 and 248 bp) and a small (51-bp) intron. The gene encodes the major β-carboxyl esterase (*EST-6*) that is transferred by *D. melanogaster* males to females in the seminal fluid during copulation (Richmond *et al.* 1980) and affects the female’s consequent behavior and mating proclivity (Gromko *et al.* 1984). The *Est-6* gene is duplicated (Collet *et al.* 1990) but there is evidence that the adjacent genetic duplicate, referred to as *Est-P* (Collet *et al.* 1990) or *Est-7* (Dumancic *et al.* 1997), may be a pseudogene (ψ*Est-6*, Balakirev and Ayala 1996; Balakirev *et al.* 2003; but see Dumancic *et al.* 1997). The β-esterase gene cluster in other *Drosophila* species also includes two (three in *D. pseudoobscura*) closely linked genes (Yenikolopov *et al.* 1989; Brady *et al.* 1990; East *et al.* 1990; Oakeshott *et al.* 1993, 1995).

The expression of *Est-6* in *D. melanogaster* has been investigated using *P* element-mediated transformation...
(Ludwig et al. 1993; Healy et al. 1996; Tamarina et al. 1997). Within the ~1.2 kb of the 5′-flanking region, several independently acting cis-regulatory promoter elements that control the expression of the gene in different tissues have been identified. Game and Oakeshott (1990) investigated restriction site polymorphism and its association with functional variation within a 21.5-kb region including the Est-6 gene and found that a restriction polymorphism at an Rsal site in the 5′-flanking region of Est-6 shows a significant association with male amount and activity of EST-6. Given other evidence showing that differences in male EST-6 activity affect the reproductive success of their mates (Richmond et al. 1990), Game and Oakeshott (1990) concluded that Est-6 cis-acting regulatory polymorphisms may be important contributors to adaptive variation. Indeed, Oakeshott et al. (1994) and Saad et al. (1994) have detected significant associations between the fitness components (preadult viability, development time, time to mating, remating frequency, egg production, and fertility) of D. melanogaster and the EST-6 activity level.

Cooke and Oakeshott (1989) sequenced the complete coding region of Est-6 in 13 D. melanogaster lines in an Australian population (chosen so as to include all allozyme variants known in the population). They suggested that the main Fast and Slow allozymes differ by two amino acids (Asn/Asp at position 237 and Thr/Ala at position 247; but see Hasson and Eanes 1996 and Balakirev et al. 1999) and considered these two amino acid replacements as the most likely targets for selection underlying the previously detected latitudinal clines (Oakeshott et al. 1981). Odgers et al. (1995) sequenced 974 bp of the Est-6 5′-flanking region in D. melanogaster and identified a nucleotide substitution responsible for the Rsal polymorphism (T → G at −531). They also revealed the presence of two highly diverged haplotype groups and a peak of polymorphism around the Rsal site. Odgers et al. (1995) showed that the Rsal+ haplotype group yields ~25% more EST-6 enzyme activity in adult males than does the Rsal− haplotype and detected weak disequilibrium between the promoter polymorphism and the Fast/Slow allozyme polymorphism. Later, Odgers et al. (2002) carried out Pelement-mediated germ-line transformation, fusing representative promoter alleles to an identical Est-6 coding region. They found a twofold difference in EST-6 activity in the male anterior sperm ejaculatory duct. Odgers et al. (2002) also conducted restriction fragment length polymorphism (RFLP) and sequencing of the promoter region in populations from Africa, America, Asia, and Australia and detected significant deviation from neutral expectations in the non-African samples but not in the African one. Hasson and Eanes (1996) investigated the nucleotide polymorphism of the Est-6 coding region in 16 lines from disparate parts of the world, selected on the basis of the presence/absence of the cosmopolitan inversion In (3L) Payne, and detected shared polymorphisms between St and In (3L) Payne chromosomes, indicating extensive genetic exchange between arrangements. Balakirev et al. (1999) sequenced 15 alleles of the Est-6 coding region from a Californian population and found two highly differentiated haplotypes, one encompassing the Fast alleles and the other consisting of Slow alleles. They also detected a distinct peak of increased variation in the region surrounding the replacement site responsible for the EST-6 Fast/Slow allozyme polymorphism and suggested that balancing selection might be involved in the polymorphism. All these studies involve samples that are too small (Cooke and Oakeshott 1989; Hasson and Eanes 1996; Balakirev et al. 1999), nonrandom, or both (Cooke and Oakeshott 1989; Hasson and Eanes 1996) and thus unsuitable for certain population genetic tests (Hudson et al. 1994; Simonsen et al. 1995).

We (Balakirev et al. 2002) increased the sample size and the length of the region sequenced to carry out significant tests of neutrality and to analyze the possible association between the regulatory and structural nucleotide polymorphism, seeking also to test for linkage disequilibrium within the gene region, a possibility suggested by the patterns observed in our previous study (Balakirev et al. 1999). We investigated the 5′-flanking, coding, and 3′-flanking regions of the Est-6 gene (3062 bp total) in a random sample of 30 lines (and thus large enough for the population genetic tests; see Hudson et al. 1994; Simonsen et al. 1995) of D. melanogaster from a natural population of California. We detected a highly structured pattern of variability, with distinctive features in the coding and 5′-flanking regions. We discovered two distinct allelic lineages for the promoter and coding region of the Est-6 gene. The pattern of variability was complex and differed between the coding and the 5′-flanking regions, although the level of nucleotide diversity was very similar in the two regions. We detected strong linkage disequilibrium within the 5′-flanking region and Est-6 coding region separately but it was much less pronounced between these two functional regions of the gene. The neutrality tests of Kelly (1997) and Wall (1999) incorporating recombination were highly significant for the studied regions. We suggested that the Est-6 nucleotide polymorphism is shaped by a combination of directional and balancing selection acting on the promoter and coding region polymorphisms and by interactions between the two regions due to different degrees of hitchhiking (Balakirev et al. 2002).

We now present the analysis of nucleotide variation of the Est-6 gene region in three additional samples of D. melanogaster derived from the natural populations of East Africa (Zimbabwe), Europe (Spain), and South America (Venezuela). The motivation for examining this gene in different populations is to analyze the pattern of nucleotide variation in the ancestral (African) and derived (European and American) D. melanogaster populations; we attempt further to clarify the question
concerning the evolutionary forces shaping the regulatory (RsaI+/RsaI−) and structural [Fast/Slow (F/S)] nucleotide polymorphisms. Odgers et al. (1995, 2002) could not analyze the association between the regulatory and structural nucleotide polymorphisms, because they did not sequence the Est-6 coding region in the same lines of D. melanogaster for which they obtained the promoter region sequences.

MATERIALS AND METHODS

Drosophila strains: D. melanogaster strains were derived from random samples of wild flies collected in Europe (Spain), North America (California), and South America (Venezuela). The strains were made fully homozygous for the third chromosome by crosses with balancer stocks, as described by Seager and Ayala (1982). The strains were named, in accordance with the electrophoretic alleles they carry for esterase-6 (the letter before the hyphen) and superoxide dismutase (the letter after the number), Ultra Slow (US), Slow (S), and Fast (F) (Figure 1). Chung-I Wu kindly provided the D. melanogaster strains from East Africa (Sengwa and Harare, Zimbabwe). The strain Zin S-4F (Zimbabwe) is from F. J. Ayala’s laboratory.

DNA extraction, amplification, and sequencing: Methods are as previously described (Balakirev et al. 2003). The sequences of both strands were determined for each line, using 12 overlapping internal primers spaced, on average, 350 nucleotides. (See GenBank accessions AF526538–AF526559, AF150809–AF150815, AF147095–147102, and AF217624–AF217645). At least two independent PCR amplifications were sequenced for each polymorphic site in all D. melanogaster strains to prevent possible PCR and sequencing errors.

DNA sequence analysis: The Est-6 sequences were assembled using the program SeqMan (Lasergene, 1994–1997; DNASTAR, Madison, WI). The computer programs DnaSP, version 3.4 (Rozas and Rozas 1999), and PROSEQ, version 2.4 (Filatov and Charlesworth 1999), were used for most intraspecific analyses. Departures from neutral expectations were investigated using Kelly’s (1997) and Wall’s (1999) tests on the basis of linkage disequilibrium between segregating sites and incorporating recombination. The permutation approach of Hudson et al. (1992a,b) was used to estimate the significance of sequence differences between populations and haplotype families. Simulations based on the algorithms of the coalescent process with recombination (Hudson 1990) were performed with the PROSEQ program to estimate the probabilities of the observed values of Kelly’s ZS and Wall’s B and Q statistics. The coalescent approach was also used to estimate confidence intervals for the nucleotide diversity values. The program Geneconv version 1.81 (Sawyer 1999) was used to detect gene conversion events. The population recombination rate was analyzed with the permutation-based approach (McVean et al. 2002) on the basis of the approximate-likelihood coalescent method of Hudson (2001).

RESULTS

Nucleotide polymorphism and recombination: The sequenced region consists of 3066 bp (2498 bp in the African sample). Figure 1 shows a total of 121 polymorphic sites (124 mutations because of three different nucleotides at each of positions 763, 1391, and 2396) in a sample of 78 sequences of the Est-6 gene from four populations of D. melanogaster: 45 sites (46 mutations) in the 5'-flanking region (3 sites, positions 329, 405, and 424, are associated with deletions), 49 sites (51 mutations) in exon I, 2 sites in the intron, 5 sites in exon II, and 20 sites in the intergenic region. Within the Est-6 exons we detected 20 replacement and 34 synonymous polymorphic sites. Nine length polymorphisms, six deletions (▲1–▲6), and three insertions (▼1–▼3) occur within the whole sequenced region (Figure 1).

The length of the 5'-flanking region sequenced in the East-African sample is 619 bp but 1183 bp in the other samples. To obtain comparable estimates of nucleotide variation in all samples, we restrict the analysis of the 5'-flanking region to the 619 bp (“standard length”) sequenced in all. Table 1 shows estimates of nucleotide diversity for the standard length of the Est-6 gene and flanking regions. The π value for the full sequence is 0.0060 ± 0.0005, which is within the range of values observed in other highly recombinating gene regions of D. melanogaster (Moriyama and Powell 1996). The π value is very similar in the 5'-flanking (0.0060 ± 0.0007) and Est-6 regions (0.0057 ± 0.0005), but higher in the intergenic region (0.0094 ± 0.0018). The synonymous variation (0.0160) is 6.7 times higher than the nonsynonymous variation (0.0024) in the Est-6 coding region. This sort of difference is expected if there is selective constraint on the Est-6 nonsynonymous substitution rate. The level of silent divergence is at least 2.0 times higher for the Est-6 gene than for the 5'-flanking or intergenic region (Table 1). The level of nucleotide diversity is highest in the African sample (π = 0.0092 ± 0.0008) and lowest in the sample from South America (π = 0.0034 ± 0.0007). Intermediate (and very similar) values of nucleotide diversity are observed in the European (π = 0.0055 ± 0.0008) and North American (π = 0.0060 ± 0.0008) samples (Table 1).

Previously, we detected in the California population lower polymorphism in the coding region of the S haplotypes than in that of the F haplotypes and lower polymorphism in the promoter region of the Rsd− haplotypes than in that of the Rsd+/Rsd− haplotypes. We also noted that the “double sweep” (Rsd−/S) haplotypes (the haplotypes that have the most common mutations in both the promoter and coding region) were least variable (Balakirev et al. 2002). A similar tendency is observed in the East-African and European samples but not in the South American sample (Table 2). The South American sample is unique in the sense that it has no F allelic lineage at Est-6 (see Figure 1). [We note that this population also lacks the S allele at the Sod locus (Hudson et al. 1994).]

The method of Hudson and Kaplan (1985) reveals a minimum of 20 recombination events in the whole region analyzed: 3 for the 5'-flanking region, 16 for the Est-6 gene, and 1 between them. The population recombination rate (McVean et al. 2002) is 0.0216 for
Figure 1.—The lines of *D. melanogaster* from East Africa [Zimbabwe (Zim)], Europe [Barcelona, Spain (Bar)], North America [El Rio, California (ER)], and South America [Caracas, Venezuela (Ven)] are presented sequentially. The lines within each population are grouped according to their genetic similarity. The S, US, and F letters before the line numbers refer to the EST-6 allozymes, Slow, Ultra-Slow, and Fast. The S and F after the numbers refer to the allozyme polymorphism at the *Sod* locus (except in Zimbabwe, where *Sod* has not been investigated) and have been previously used to tag these lines. The second letter in the Zimbabwe lines (not in Zim S-44F) refers to the locality of collection (Sengwa or Harare). The line Zim S-44F is from F. J. Ayala's laboratory. The numbers above the top sequence represent the position of segregating sites and the start of a deletion or insertion. Nucleotides are numbered from the beginning of our sequence (position 32 in Collet et al. 1990). The coding region (exon I and exon II) of the *Est-6* gene is underlined below the reference sequence (ER S-26F). Amino acid replacement polymorphisms are marked with asterisks. The *Rsa* I polymorphism is determined by site 653, where *Rsa* I/H11001 has T and *Rsa* I/H11002 has G; the S-F allozyme polymorphism is determined by site 1959, where S has A (asparagine) and F has G (aspartic acid). Dots indicate the same nucleotide as the reference sequence. Hyphens represent deleted nucleotides. Question marks indicate missing data. /H17009 denotes a deletion; † denotes the absence of a deletion; /H17010 denotes an insertion; ‡ denotes the absence of an insertion. 1, 5-bp deletion of CTTTT; 2, 19-bp deletion of TTCTATTTGTCGCAAGCA; 3, single-nucleotide deletion of T; 4, 2-bp deletion of AA; 5, 9-bp deletion of GAACT; 6, 3-bp deletion of GAT; 7, 3-bp insertion of TGT.
1905

Figure 1.—Continued.
### TABLE 1

Nucleotide diversity and divergence in the *Est-6* gene region of *D. melanogaster*

<table>
<thead>
<tr>
<th></th>
<th>5'-flanking region (619 bp)</th>
<th>Syn (416 bp)</th>
<th>Nsyn (1274 bp)</th>
<th>Total (1686 bp)</th>
<th>Intergenic region (193 bp)</th>
<th>Silent sites (1228 bp)</th>
<th>All sites (2498 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>π</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0060</td>
<td>0.0160</td>
<td>0.0024</td>
<td>0.0057</td>
<td>0.0094</td>
<td>0.0097</td>
<td>0.0060</td>
</tr>
<tr>
<td><strong>θ</strong></td>
<td>0.0084</td>
<td>0.0187</td>
<td>0.0032</td>
<td>0.0066</td>
<td>0.0210</td>
<td>0.0137</td>
<td>0.0082</td>
</tr>
<tr>
<td><strong>K</strong></td>
<td>0.0514</td>
<td>0.1474</td>
<td>0.0213</td>
<td>0.0495</td>
<td>0.0707</td>
<td>0.0847</td>
<td>0.0519</td>
</tr>
</tbody>
</table>

Total (78 lines)

| **π**    | 0.0126                      | 0.0194       | 0.0035         | 0.0073          | 0.0148                   | 0.0150                 | 0.0092             |
| **θ**    | 0.0127                      | 0.0256       | 0.0037         | 0.0085          | 0.0189                   | 0.0173                 | 0.0103             |
| **K**    | 0.0539                      | 0.1454       | 0.0224         | 0.0499          | 0.0722                   | 0.0856                 | 0.0529             |

Zim (12 lines)

| **π**    | 0.0043                      | 0.0169       | 0.0023         | 0.0058          | 0.0062                   | 0.0087                 | 0.0055             |
| **θ**    | 0.0053                      | 0.0146       | 0.0019         | 0.0048          | 0.0121                   | 0.0082                 | 0.0050             |
| **K**    | 0.0504                      | 0.1476       | 0.0213         | 0.0494          | 0.0709                   | 0.0838                 | 0.0516             |

Bar (18 lines)

| **π**    | 0.0044                      | 0.0152       | 0.0026         | 0.0057          | 0.0141                   | 0.0094                 | 0.0060             |
| **θ**    | 0.0053                      | 0.0156       | 0.0027         | 0.0056          | 0.0186                   | 0.0095                 | 0.0061             |
| **K**    | 0.0500                      | 0.1469       | 0.0215         | 0.0495          | 0.0716                   | 0.0836                 | 0.0516             |

ER (28 lines)

| **π**    | 0.0044                      | 0.0108       | 0.0013         | 0.0035          | 0               | 0.0056                 | 0.0034             |
| **θ**    | 0.0042                      | 0.0089       | 0.0016         | 0.0033          | 0               | 0.0050                 | 0.0033             |
| **K**    | 0.0501                      | 0.1493       | 0.0206         | 0.0495          | 0.0682                   | 0.0841                 | 0.0514             |

Ven (20 lines)

bp, the number of sites; **π**, the average number of nucleotide differences per site among all pairs of sequences (Nei 1987, p. 256); **θ**, 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); **K**, the average proportion of nucleotide differences between *D. melanogaster* and *D. simulans*, corrected according to Jukes and Cantor (1969); Syn, synonymous; Nsyn, nonsynonymous. The segregating sites associated with indels are excluded from the **π**, **θ**, and **K** calculations. The 5'-flanking region is restricted to the 619 bp sequenced in all samples. Zim (Zimbabwe), Bar (Barcelona), ER (El Rio, California), and Ven (Venezuela) refer to the sampled localities.

The method of Sawyer (1989, 1999) detects gene conversion events within the *Est-6* gene in all samples except Venezuela. The number of significant fragments varies from 1 (Africa) to 14 (North America). The average length of fragments is 636 bp (range 314–1183 bp). The conversion events are less pronounced in the protein alignment (only 2 significant fragments, 1 in Africa and 1 in North America), which suggests the involvement mostly of silent sites in significant fragments of the nucleotide alignment.

**Haplotype structure and differentiation of populations:** Maximum haplotype diversity occurs in East Africa (*H_{div} = 1.000;* no identical sequence pairs); less occurs in Europe (*H_{div} = 0.895;* 16 identical sequence pairs) and North America (*H_{div} = 0.947;* 20 identical sequence pairs); and the minimum occurs in South America (*H_{div} = 0.621;* 72 identical sequence pairs).

Figure 2 shows a neighbor-joining tree of the *Est-6* sequences (standard length). Due to recombination and gene conversion, this tree is not a good reflection of the genealogical process, but it serves to show the genetic structure of the data. The tree shows a relative absence of geographic structure: the sequences from a given population do not all group together. However, recombination has not completely erased all information, since there are two clusters of haplotypes related to *Rsa* polymorphism (data not shown). The first cluster includes the sequences with the *Rsa*- haplotypes (all
strains from Ven S-10F at the top to ER F-1461S at the bottom); the second cluster contains the Rsal+ haplotypes (all strains from ER S-255S down to Ven S-2F). The Rsal−/Rsal+ clusters are even more apparent in the tree for the promoter region only (data not shown).

If we restrict the analysis only to the coding region, the two clusters obtained differ to some extent (but not exclusively) with respect to the S and F haplotypes (data not shown).

Odojers et al. (1995) described two groups of haplotypes for the 5′-flanking region of the Est-6 gene of D. melanogaster from Australia. We detected two groups of haplotypes both for the Est-6 gene (including the 5′-flanking region) and for the ψEst-6 putative pseudogene from North America (California; Balakirev and Ayala 1996; Balakirev et al. 1999, 2002, 2003). Two significantly divergent sequence types are also detected in South America (Figure 3A), where only the Slow Est-6 allozyme occurs. The average number of nucleotide differences (K) between the two haplotypes is 11.286. This is comparable with the differences between Rsal+/Rsal− (K = 6,720) and F/S (K = 11,890) allelic lineages in California (Balakirev et al. 2002). The permutation test (Hudson et al. 1992a) of the Venezuelan haplotypes is highly significant, Κ^{*} = 0.5867 (P = 0.0000). Sequence dimorphism is less pronounced in the European sample (Figure 3B). The two divergent sequence types are not associated with Est-6 allozyme variation (South America) or imperfectly associated (Europe, North America). The African sample (Figure 3C) has no clear sequence dimorphism (although all S haplotypes but one cluster together).

The estimates of population differentiation (Hudson et al. 1992a) are fairly similar between the pairs Zim-Bar (FST = 0.0653), Zim-ER (FST = 0.0398), Bar-Ven (FST = 0.1093), and ER-Ven (FST = 0.0920) (for locality abbreviations see Figure 1). The maximal and minimal FST values are obtained, respectively, for the pairs Zim-Ven (FST =

**TABLE 2**

Nucleotide diversity in different allelic lineages of the Est-6 gene

<table>
<thead>
<tr>
<th>Population</th>
<th>Rsal−</th>
<th>Rsal+</th>
<th>S</th>
<th>F</th>
<th>Double-sweep haplotypes</th>
<th>Other haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zim</td>
<td>N = 4, S = 4</td>
<td>N = 8, S = 19</td>
<td>N = 7, S = 25</td>
<td>N = 5, S = 28</td>
<td>N = 3, S = 16</td>
<td>N = 9, S = 73</td>
</tr>
<tr>
<td></td>
<td>π = 0.0033**+</td>
<td>π = 0.0123**+</td>
<td>π = 0.0053</td>
<td>π = 0.0077</td>
<td>π = 0.0043**</td>
<td>π = 0.0099***</td>
</tr>
<tr>
<td></td>
<td>θ = 0.0036**+</td>
<td>θ = 0.0122**+</td>
<td>θ = 0.0061</td>
<td>θ = 0.0080</td>
<td>θ = 0.0043**</td>
<td>θ = 0.0108***</td>
</tr>
<tr>
<td></td>
<td>D = −0.7801</td>
<td>D = −0.2164</td>
<td>D = −0.6953</td>
<td>D = −0.2444</td>
<td>D = NA</td>
<td>D = −0.4834</td>
</tr>
</tbody>
</table>

| Bar        | N = 14, S = 10 | N = 4, S = 11 | N = 13, S = 14 | N = 5, S = 12 | N = 12, S = 11 | N = 6, S = 33 |
|            | π = 0.0022 | π = 0.0048 | π = 0.0027 | π = 0.0036 | π = 0.0020**+†† | π = 0.0060* |
|            | θ = 0.0028 | θ = 0.0053 | θ = 0.0027 | θ = 0.0034 | θ = 0.0015***+†† | θ = 0.0058***+† |
|            | D = −0.7232 | D = −0.8374 | D = 0.0484 | D = 0.3014 | D = 1.4468** | D = 0.2113 |

| ER         | N = 19, S = 10 | N = 9, S = 14 | N = 19, S = 16 | N = 9, S = 30 | N = 13, S = 16 | N = 15, S = 52 |
|            | π = 0.0012 | π = 0.0050** | π = 0.0021***+† | π = 0.0070* | π = 0.0013***+†† | π = 0.0079***+† |
|            | θ = 0.0026* | θ = 0.0050 | θ = 0.0027***+† | θ = 0.0066* | θ = 0.0021***+†† | θ = 0.0064***† |
|            | D = −1.8648***+† | D = 0.3769 | D = −0.9273 | D = 0.3613 | D = −1.5061** | D = 1.0034* |

| Ven        | N = 15, S = 11 | N = 5, S = 1 | N = 20, S = 20 | N = 0 | N = 15, S = 15 | N = 5, S = 1 |
|            | π = 0.0016***+† | π = 0.0004***+† | π = 0.0035 | π = 0.0009** | π = 0.0002 | π = 0.0061***† |
|            | θ = 0.0050***+†† | θ = 0.0004***+†† | θ = 0.0033 | θ = 0.0019***+† | θ = 0.0002***+† | θ = 0.0028***† |
|            | D = −1.9446***+† | D = 0.3769 | D = −0.8345 | D = 0.3434 | D = −2.0801***+†† | D = −0.8165 |

| Bar + ER   | N = 33, S = 15 | N = 13, S = 14 | N = 32, S = 22 | N = 14, S = 30 | N = 25, S = 17 | N = 21, S = 55 |
|            | π = 0.0017 | π = 0.0046* | π = 0.0024* | π = 0.0059* | π = 0.0017***+†† | π = 0.0075***+†† |
|            | θ = 0.0033 | θ = 0.0040 | θ = 0.0032* | θ = 0.0056 | θ = 0.0018***+†† | θ = 0.0061***†|
|            | D = −1.591***+† | D = 0.5673 | D = −0.8897 | D = 0.2476 | D = −0.1648 | D = 0.8988 |

| Zim + Bar + ER | N = 37, S = 7 | N = 21, S = 19 | N = 39, S = 36 | N = 19, S = 41 | N = 28, S = 26 | N = 30, S = 88 |
|               | π = 0.0011***+† | π = 0.0067***+† | π = 0.0030 | π = 0.0070* | π = 0.0020***+†† | π = 0.0086***+†† |
|               | θ = 0.0027***+†† | θ = 0.0051 | θ = 0.0088***+† | θ = 0.0070 | θ = 0.0027***+†† | θ = 0.0090***+†† |
|               | D = −1.7418***+† | D = −0.8864 | D = −1.4082* | D = −0.0039 | D = −0.9292* | D = −0.1673 |

N, number of haplotypes; S, number of polymorphic sites. The data for the 5′-flanking region are based on 1183 bp, except for Zimbabwe and Zim + Bar + ER, which include only 619 bp. The haplotype analyses are based on the "standard" length (see Results). The π and θ values are significantly lower or higher than the corresponding values in the counterpart group of haplotypes with the population recombination rate equal to 0.0216 (*P < 0.05; **P < 0.01; ***P < 0.001) or without recombination (P < 0.05; **P < 0.01; ***P < 0.001). Tajima’s (1989) D is significant by the coalescent simulations with the population recombination rate equal to 0.0216 (*P < 0.05; **P < 0.01; ***P < 0.001) or without recombination (P < 0.05; **P < 0.01; ***P < 0.001).
and 1800 (Figure 4). The first valley region includes Linkage disequilibrium is notably low in the African population (from which the others derive by migration). Fisher’s criterion corroborates our hypothesis (Oakeshott et al. 1993, 1995). A third valley region encompasses the potential N-linked glycosylation site, corresponding to codon position 1258–1260 (1877–1879 in our coordinates). The correspondence between the level of polymorphism and localities of functionally important sites implicated in the catalytic mechanism suggests that the observed valley regions reflect functional constraint.

We have measured heterogeneity in the distribution of silent polymorphic sites along the Est-6 sequence and discordance between the level of within-melanogaster polymorphism and the melanogaster-simulans divergence by means of Goss and Lewontin’s (1996) and McDonald’s (1996, 1998) statistics and have assessed their significance by Monte Carlo simulations of the coalescent model incorporating recombination (McDonald 1996, 1998). On the basis of 10,000 simulations, with the recombination parameters varying from 1 to 64, the tests are not significant for any of the separate samples or for the combined data set (data not shown).

**Linkage disequilibrium:** Linkage disequilibrium (LD) is measured by calculating the P value of Fisher’s exact test in all pairwise comparisons between polymorphic sites. For the whole standard region (2498 bp) there are 1485 pairwise comparisons and 467 (31.45%) of them are significant. (With the Bonferroni correction, 11.92% remain significant; Bonferroni-corrected values are italicized in the ensuing sentences.) For the 5’-flanking region 25 of 78 (32.05%; 23.08%) pairwise comparisons are significant. For the Est-6 coding region (including the intron) 219 of 528 (41.48%; 23.11%) comparisons are significant. There are 19.58% (11.17%) significant associations between the 5’-flanking region and the Est-6 gene coding region. The proportion of pairs of sites with LD values significantly different from zero, at the 5% level, is much higher within the 5’-flanking region and Est-6 coding region (244 of 606 pairwise comparisons) than between them (84 of 429, Fisher’s exact test, P < 0.001; Fisher’s criterion = 52.919; P < 0.001). This observation corroborates our hypothesis (Balakirev et al. 2002) that the promoter and coded regions are subject to separate selection processes.

Linkage disequilibrium is notably low in the African sample: only 1.23% significant associations are in this vicinity (11.92% in the combined sample). Another way to address contagious regions is by means of Sliding-window analysis: Figure 4 shows the distribution of polymorphism along the Est-6 sequences. There is a distinct peak in the 5’ flanking region, which includes the Rsal+/Rsal− site (position 653 in Figure 1). Oders et al. (1995) detected this peak of variation in an Australian population of D. melanogaster. We also detected this peak (Balakirev et al. 1999, 2002, 2003) in the Californian population of D. melanogaster. Another peak of variation occurs around the F/S site except in Venezuela. We detected this peak (Balakirev et al. 1999, 2002, 2003) in our Californian data and also in data of Hasson and Eanes (1996) and Cooke and Oakeshott (1989) and suggested that it may reflect the effect of balancing selection (Strobeck 1983; Hudson and Kaplan 1988) between the F and S haplotypes, rather than within them (Balakirev et al. 2002). The absence of the peak in Venezuela may be a consequence of the absence of F haplotypes in this sample (Figure 1). The strong presence of both promoter (Rsal+/Rsal−) and coding region (F/S) peaks in the African sample (Figure 4) suggests that these polymorphic sites were targets of balancing selection already in the African population (from which the others derive by migration).

The valley regions located between the peaks of nucleotide variation are centered around positions 350, 1200, and 1800 (Figure 4). The first valley region includes nearly 400 bp upstream of the Est-6 coding region. Kar-
sample, but 22.59, 21.45, and 37.68% are in the European, North American, and South American samples, respectively. Figure 5 shows the distribution of $D$ values along the whole region studied. A notable peak is around the F/S site and a less pronounced peak is around the Rsal−/Rsal+ site.

The significance of Pearson’s correlation coefficient between LD and physical distance between sites is estimated by 10,000 permutations (McVean et al. 2002). For all samples, except South America, there is significant decline in LD with increasing distance (Table 4). The strong haplotype structure and pattern of linkage disequilibrium suggest that the South American population originated from a recent admixture of genetically differentiated populations.

**Tests of neutrality:** The tests of Hudson et al. (1987), Tajima (1989), and Depaulis and Veuille (1998) do not reveal any significant deviation from neutrality for the Est-6 gene region in any of the four populations of *D. melanogaster* (see also Balakirev et al. 2002). However, Kelly’s (1997) $Z_{ss}$ and Wall’s (1999) $B$ and $Q$ tests detect significant deviations from neutrality in the non-African samples, with the population recombination rate ranging from 0.005 to 0.010 (Table 5; data for $B$ and $Q$ are not shown). The tests fail to detect any significant deviation from neutrality for the African sample, even when using 0.0664 as the recombination rate (laboratory estimate and based on the physical and genetic maps of *D. melanogaster*; J. M. Comeron, personal communication; Comeron et al. 1999; Balakirev et al. 2002), which is at least 2.5 times higher than the value of recombination obtained by the method of McVean et al. (2002) (Table 3). The significant values of Kelly’s and Wall’s statistics are grouped around the peaks of linkage disequilibrium and centered around the functionally important sites within both the 5’-flanking region (Rsal site) and the coding region (F/S polymorphism) of the Est-6 gene (data not shown), which has been interpreted as evidence that these sites are targets of balancing selection (Ayala et al. 2002; Balakirev et al. 2002, 2003).

**DISCUSSION**

We have investigated nucleotide polymorphism in the Est-6 gene region in four populations of *D. melanogaster* from Zimbabwe, Spain, California, and Venezuela. A dimorphic haplotype structure exists in the North American sample, which is not perfectly associated with the Est-6 allozyme variation (S/F) and in South America, where there are no Est-6 F haplotypes. The presence

**Figure 2.** Neighbor-joining tree of the Est-6 haplotypes of *D. melanogaster*, based on Kimura’s two-parameter distance. The numbers are bootstrap probability values based on 10,000 replications. The trees are based on the Est-6 standard length.
of two or more highly diverged haplotypes has been interpreted as a result of positive selection in *D. melanogaster* (see, e.g., Hudson et al. 1994, 1997; Bénassi et al. 1999; Labate et al. 1999). Teeter et al. (2000) investigated single-nucleotide polymorphism in 66 sequences of *D. melanogaster* spaced at 5- to 20-cM intervals and generated a map with no gaps greater than one-half of a chromosome arm (Teeter et al. 2000). Two-thirds of all sequences were dimorphic. If the dimorphism results from positive selection, Teeter et al. (2000) estimate that one site for every few kilobases would be subject to strong positive selection, which seems improbable. Teeter et al. (2000) suggest that admixture between two differentiated populations of *D. melanogaster* would account for and be a more appropriate explanation of the dimorphism. Suggestions of admixture have also been made on the basis of nucleotide sequencing, RFLP, and allozyme analyses of *D. melanogaster* populations (e.g., David and Capy 1988; Singh and Long 1992; Richter et al. 1997; Hasson et al. 1998).

Our *Est-6* data are compatible with this proposal. We have found a strong dimorphic haplotype structure in three other *D. melanogaster* genes on the third chromosome, *Sod* (Hudson et al. 1997), *tinman*, and *bagpipe* (E. S. Balakirev and F. J. Ayala, unpublished data), which may also have resulted from population admixture. Nevertheless, the *Est-6* data suggest that positive selection may also contribute to the observed patterns: from positive selection, Teeter et al. (2000) estimate that one site for every few kilobases would be subject to balanced selection would account for the elevated nucleotide variation and linkage disequilibrium around the target polymorphic sites (RsaI−/RsaI+ in the promoter region and F/S in the coding region), while directional selection would yield an excess of very similar sequences exhibiting a very low level of variability (RsaI− and S allelic lineages, in the promoter and coding region, respectively).

**Figure 3.—**Neighbor-joining tree of the *Est-6* haplotypes of *D. melanogaster* from South America (A), Europe (B), and Africa (C) based on Kimura’s two-parameter distance. The numbers are bootstrap probability values based on 10,000 replications.
The African sample has the highest level of nucleotide diversity and the lowest level of linkage disequilibrium. The non-African samples show a pattern of haplotype distribution consistent with selective sweep hypotheses in the history of the species. The distribution of haplotype frequency in non-African samples is highly asymmetric: from a total of 66 sequences, 52 belong to the S haplotype and 48 belong to Rsal− haplotype. The haplotype test (Hudson et al. 1994) is significant for the North and South American (excluding the recombinant strain Ven S-13F) samples, but not significant for the European sample. We conclude that bottlenecks have been an important evolutionary factor changing the genetic composition of colonizing D. melanogaster populations. The haplotype structure and polymorphism of the Est-6 gene region are in accordance with the general pattern of relationships between the African and non-African populations of D. melanogaster (Andolfatto 2001; Aquadro et al. 2001). However, the peaks of nucleotide variation in the African sample, centered on functionally important sites (Figure 4), suggest that this population is not in mutation-drift equilibrium. The footprints of directional selection have been previously shown in African populations (e.g., Mousset et al. 2003).

We found lower polymorphism in the S than in the F haplotypes (coding region) and lower polymorphism in the Rsal− than in the Rsal+ haplotypes (promoter region) in the California population (Balakirev et al. 2002). The same pattern occurs in the other populations (excluding Venezuela, where no F haplotypes occur), as well as in the total data set encompassing all four populations (Table 2): π is six times higher for the Rsal+ than for the Rsal− haplotypes; for the coding region, π is twice as large for the F as for the S haplotypes but double (0.00695) for the F haplotypes. Thus the
TABLE 4
Correlation between linkage disequilibrium and physical distance between the *Est-6* (full-sequence) polymorphic sites

<table>
<thead>
<tr>
<th>Population</th>
<th>$LD_{Fisher}$</th>
<th>$r^2$</th>
<th>$D'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zim</td>
<td>0.0123</td>
<td>-0.0928 (P = 0.0000)</td>
<td>-0.0507 (P = 0.0383)</td>
</tr>
<tr>
<td>Bar</td>
<td>0.2259</td>
<td>-0.2918 (P = 0.0000)</td>
<td>-0.1328 (P = 0.0108)</td>
</tr>
<tr>
<td>ER</td>
<td>0.2145</td>
<td>-0.1525 (P = 0.0004)</td>
<td>-0.1963 (P = 0.0000)</td>
</tr>
<tr>
<td>Ven</td>
<td>0.3768</td>
<td>+0.0626 (P = 0.9333)</td>
<td>+0.0443 (P = 0.8203)</td>
</tr>
<tr>
<td>All</td>
<td>0.1293</td>
<td>-0.1419 (P = 0.0000)</td>
<td>-0.0603 (P = 0.0091)</td>
</tr>
</tbody>
</table>

$LD_{Fisher}$ is the proportion of significant linkage disequilibrium revealed by Fisher’s exact test using all polymorphic sites (including singletons).

lower variability among *Rsa*− and S haplotypes is not limited to the California population. But the differences are smaller in the African sample, which could indicate that the *Rsa*− and S haplotypes increased in frequency in Europe and America after their colonization.

We propose that the *Rsa*+/F (zero-sweep) haplotypes may represent the ancestral condition (Balakirev et al. 2002). The frequency of these haplotypes is higher in Africa (0.333) than elsewhere (0.091). We also suggest that the *Rsa*−/S (double-sweep) haplotypes have evolved under directional selection, since they are less variable but more frequent in non-African samples (0.606) than in African (0.250). Directional selection, however, does not lead toward fixation of the double-sweep haplotypes in the derived populations because of balancing selection maintaining both divergent haplotypes (*Rsa*−/*Rsa*+ and F/S) in the promoter and coding regions (Balakirev et al. 2002).

The population data available suggest two different migrations of *D. melanogaster* during the expansion period from the African continent: (1) Africa → Europe → North America and (2) Africa → South America (see also David and Capy 1988; Singh and Long 1992). The second migration is supported by the fact that the East-African and South American samples share a deletion (●6, Figure 1) that is absent in other samples. This deletion is present in 5 of 12 East-African strains but absent in Europe and North America (Figure 1). Gaps constitute a valuable source of phylogenetic information (Griebel and Wheeler 1999). The absence of the F *Est-6* allele (and of the S *Sod* allele; Hudson et al. 1994) also suggests that the South American population does not derive from Europe or America. The South American population might represent an admixture of migrants from North America and Africa. The most common haplotype (*Rsa*−/S) is from North America, while the haplotype *Rsa*+/S clusters with most of the African strains (Figure 2). The admixture would have been recent, since the strong haplotype structure has not been eroded by recombination (linkage disequilibrium is highest in the South America sample).

We are grateful to G. McVean, D. A. Filatov, J. K. Kelly, J. H. McDonald, J. D. Wall, J. M. Comeron, F. Depaulis, and J. Rozas for useful advice on analyses and for providing computer programs. We thank Elena Balakireva, Andrei Tatarenkov, Victor DeFilippis, Martina Zurovkova, and Carlos Márquez for encouragement and help; and

### TABLE 5
Kelly’s (1977) test of neutrality for the *Est-6* gene region

<table>
<thead>
<tr>
<th>5′-flanking region (0.6 kb)</th>
<th><em>Est-6</em> region (1.6 kb)</th>
<th>Entire region (2.4 kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zns</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.079</td>
<td>0.030</td>
</tr>
<tr>
<td>All populations</td>
<td>0.062</td>
<td>0.025</td>
</tr>
<tr>
<td>Zim</td>
<td>0.138</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>0.115</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>0.111</td>
<td>NS</td>
</tr>
<tr>
<td>Bar</td>
<td>0.540</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>0.285</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>0.227</td>
<td>0.010</td>
</tr>
<tr>
<td>ER</td>
<td>0.458</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>0.204</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>0.181</td>
<td>0.010</td>
</tr>
<tr>
<td>Ven</td>
<td>0.453</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>0.386</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>0.417</td>
<td>0.005</td>
</tr>
</tbody>
</table>

The columns give Kelly’s (1977) $Z_{ns}$ values and the recombination rates at which the test becomes significant at the 5% (0.05) or 1% (0.01) level of significance. NS, not significant, even with the laboratory-estimated recombination rate (0.0664). Zim (Zimbabwe), Bar (Barcelona), ER (El Rio, California), and Ven (Venezuela) refer to the sampled localities.
LITERATURE CITED


Communicating editor: M. Asmussen