UC Irvine

UC Irvine Previously Published Works

Title

Molecular evolution of the Est-6 gene in *Drosophila melanogaster*: contrasting patterns of DNA variability in adjacent functional regions.

Permalink

<https://escholarship.org/uc/item/1862w6rp>

Journal

Gene, 288(1-2)

ISSN

0378-1119

Authors

Balakirev, Evgeniy S
Balakirev, Elena I
Ayala, Francisco J

Publication Date

2002-04-01

DOI

10.1016/s0378-1119(02)00477-8

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: contrasting patterns of DNA variability in adjacent functional regions

Evgeniy S. Balakirev^{a,b,c}, Elena I. Balakirev^a, Francisco J. Ayala^{a,*}

^aDepartment of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525, USA

^bInstitute of Marine Biology, Vladivostok 690041, Russia

^cAcademy of Ecology, Marine Biology, and Biotechnology, Far Eastern State University, Vladivostok 690600, Russia

Received 21 August 2001; received in revised form 15 January 2002; accepted 8 February 2002

Received by T. Gojobori

Abstract

We have investigated nucleotide polymorphism at the esterase 6 gene (*Est-6*) gene, including the complete coding region (1686 bp), as well as the 5'-flanking (1183 bp) and 3'-flanking (193 bp) regions of the gene, in 30 strains of *Drosophila melanogaster* and in one strain of *Drosophila simulans*. The level of silent variation is similar in the coding and in the 3'-flanking region, but smaller in the 5'-flanking region. Strong linkage disequilibrium occurs within each region; and also, although less pronounced, between the 5'-flanking region and the rest of the gene, including the 3'-flanking region. We suggest that the pattern of nucleotide polymorphism of *Est-6* may be shaped by: (1) directional and balancing selection acting on the promoter and the coding region; and (2) interactions between the two regions that involve variable degrees of hitchhiking. The patterns of linkage disequilibrium, as well as the statistics Z_{ns} (Genetics 146 (1997) 1197) and B and Q (Genet. Res. 74 (1999) 65), may be interpreted as there being multiple targets of selection within the gene. The previously reported *Est-6* allozyme latitudinal clines may be accounted for by the interaction between selective processes in the promoter and coding regions. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: DNA polymorphism; *Drosophila melanogaster*; Esterase-6; Linkage disequilibrium; Multi-locus selection

1. Introduction

The esterase 6 (*Est-6*) gene is on the left arm of chromosome 3 of *Drosophila melanogaster*, 32.5 in the genetic map and at 68F7-F8 in the cytogenetic map. The coding region is 1686 bp long and consists of two exons (1387 and 248 bp) and a small (51 bp) intron. The gene is duplicated (Collet et al., 1990) but there is evidence that the adjacently located duplicate, referred to as *Est-P* by Collet et al. (1990) may be a pseudogene (ψ *Est-6*, Balakirev and Ayala, 1996). The EST-6 protein is transferred by *D. melanogaster* males to females in the seminal fluid during copulation and affects the female's consequent behavior and mating proclivity (Richmond et al., 1990, and references therein).

Two main allozymes (fast, F and slow, S) of EST-6 have been described, which manifest a differential response to the organophosphate inhibitor, which raises questions about the adaptive significance of the polymorphism. The main allozymes show large-scale repeatable latitudinal clines, with

the S allozyme more common at higher latitudes. This, together with other data on the temporal and geographic allozyme 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., 1995; Richmond et al., 1990).

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 F and S allozymes differed by two amino acids (Asn/Asp at position 237 and Thr/Ala at position 247) (but see Hasson and Eanes, 1996; Balakirev et al., 1999) and considered these two amino acid replacements as the most likely targets for selection underlying the latitudinal clines previously detected for the F and S allozymes. 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

Abbreviations: *Est-6*, esterase 6 gene; F, fast; S, slow; US, ultra slow

* Corresponding author. Tel.: +1-949-824-8293; fax: +1-949-824-2474.

E-mail address: fjayala@uci.edu (F.J. Ayala).

arrangements. Balakirev et al. (1999) sequenced 15 alleles of *Est-6* coding region from a Californian population and found two highly differentiated haplotypes, one encompassing the F alleles, and the other consisting of S alleles. They also detected a distinct peak of increased variation in the region surrounding the replacement site responsible for the EST-6 F/S allozyme polymorphism and suggested that balancing selection might be involved in the polymorphism. All these studies involve samples that are too small (Balakirev et al., 1999), non-random, or both (Cooke and Oakeshott, 1989; Hasson and Eanes, 1996) and thus unsuitable for many population genetic tests.

The expression of *Est-6* in *D. melanogaster* has been investigated using *P*-element-mediated transformation (Tamarina et al., 1997, and references therein). Within the ~1.2 kb of the 5'-flanking region, several independently acting *cis*-regulatory promoter elements have been identified that control the expression of the gene in different tissues. 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 *RsaI* site in the 5'-flanking region of *Est-6* shows a significant association with male amount and activity level of EST-6. Given the evidence from other studies that differences in male EST-6 activity affect the reproductive success of their mates (Richmond et al., 1990), Game and Oakeshott (1990) conclude that *Est-6 cis*-acting regulatory polymorphisms may be important contributors to adaptive variation. Indeed, significant associations have been detected between several fitness components (pre-adult viability, development time, time to mating, remating frequency, egg production, and fertility) and the EST-6 activity level.

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 *RsaI* polymorphism (T → G at -531). They also revealed the presence of two highly diverged haplotype groups and a peak of polymorphism around the *RsaI* site. By comparing their data with the results of Game and Oakeshott (1990), Odgers et al. (1995) showed that the *RsaI* + haplotype group yields ~25% more EST-6 enzyme activity in adult males than the *RsaI* - one, and detected weak disequilibrium between the promoter polymorphism and the F/S allozyme polymorphism. However, Odgers et al. (1995) did not sequence the *Est-6* coding region in the same lines of *D. melanogaster* for which they obtained the promoter region sequences, which would have allowed them to analyze the pattern and extent of the association between the regulatory and structural nucleotide polymorphism.

Balakirev et al. (1999) studied *Est-6* in 15 lines of *D. melanogaster* and found departures from neutral polymorphism, but the results were ambiguous due to the small size of the sample. We now increase sample size and the length of the region sequenced, 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 investigate 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) of *D. melanogaster* derived from a natural population of California. The detected pattern of variability is highly structured with distinctive features in the coding and 5'-flanking regions. We suggest 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 the interactions between the two regions due to different degrees of hitchhiking.

2. Materials and methods

2.1. *Drosophila* strains

The 30 *D. melanogaster* strains were derived from a random sample of wild flies collected by F.J. Ayala (October 1991) in El Rio Vineyard, Acampo, CA (USA). 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 chosen because their *Sod* gene had been previously investigated in our laboratory (Hudson et al., 1994) and named in accordance with the esterase-6 (the letter before the hyphen) and superoxide dismutase (the letter after the number) electrophoretic alleles they carry (Hudson et al., 1994), ultra slow (US), S, and F (Table 1). Fifteen lines are new. Moreover, we have sequenced the 5'-flanking region (1183 bp) for the 15 lines previously analyzed by Balakirev et al. (1999) (as well as for the new lines), so that a complete sequence of 3062 bp has been obtained for the same 30 lines.

2.2. Allozyme electrophoresis

A total of 20 flies from each line were homogenized in 20 µl 0.1 M Tris-HCl buffer, pH 8.0. The homogenates were electrophoresed for 8–9 h using Tris-borate-EDTA continuous buffer system, pH 8.6 at 100 V in an 11% starch gel. The gels were stained for EST-6 activity according to standard procedures.

2.3. DNA extraction, amplification, and sequencing

Total genomic DNA was extracted using the tissue protocol of the QIAamp Tissue Kit (QIAGEN®).

The *D. melanogaster Est-6* sequence (GenBank accessions M33780, M33781; Collet et al., 1990) was used for designing polymerase chain reaction (PCR) and sequencing primers. The primers were designed using the computer program Primer Select (Lasergene, DNASTAR, Inc., 1994–1997). The amplified fragment included 1183 bp of the 5'-flanking region, the complete *Est-6* gene (1686 bp), the intergenic region between *Est-6* and ψ *Est-6* (193 bp), and

[illegible]

127 bp of the $\psi Est-6$. The primers used for the PCR amplification reactions were: 5'-aagcttgctatatctatctgt-3' (forward primer) and 5'-catagggaatcgattcgtagctgt-3' (reverse primer). The PCR reactions were carried out in final volumes of 50 μ l using TaKaRa Ex TaqTM in accordance with the manufacturer description (Takara Biotechnology Co., Ltd.). The reaction mixtures were overlaid with mineral oil, placed in a DNA thermal cycler (Perkin Elmer Cetus), incubated 5 min at 95° and subjected to 32 cycles of denaturation, annealing, and extension: 95° for 30 s, 56° for 30 s, and 72° for 2.0 min (for the first cycle and progressively adding 3

One-tenth of each reaction volume was assayed on a 0.8% agarose gel. When the desired PCR product was detected, the remainder of the reaction was purified with Wizard™ PCR preps DNA purification system (Promega corporation). The purified PCR product was directly sequenced by the dideoxy chain-termination technique using Dye Terminator chemistry and separated with the ABI PRISM 377 automated DNA sequencer (Perkin Elmer). For each line, the sequence of both strands was determined, using 14 internal

primers spaced, on average, 350 nucleotides. (See GenBank accessions AF147095–AF147102 and AF217624–217645 for the *Est-6* sequences.) At least two independent PCR amplifications were sequenced for each polymorphic site in all *D. melanogaster* strains to prevent possible PCR or sequencing errors.

2.4. DNA sequence analysis

The *Est-6* sequences were assembled using the programs DARWIN (elaborated by Dr Robert Tyler from our laboratory) and SeqMan (Lasergene, DNASTAR, Inc., 1994–1997). Multiple alignment was carried out manually and using the program CLUSTAL W. Linkage disequilibrium between polymorphic sites was evaluated using Fisher's exact test of independence. The computer programs DnaSP, version 3.4 (Rozas and Rozas, 1999) and PROSEQ, version 2.4 (Filatov and Charlesworth, 1999) were used to analyze the data by means of the 'sliding window' method (Hudson and Kaplan, 1988), and for most intraspecific analyzes. Departures from neutral expectations were investigated using the corresponding tests (Hudson et al., 1987; Tajima, 1989; McDonald and Kreitman, 1991; Fu and Li, 1993; Hudson et al., 1994; McDonald, 1998; Kelly, 1997; Depaulis and Veuille, 1998; Wall, 1999). The non-coding sequences (intron and intergenic region) were used as a neutral reference in the HKA test. A permutation approach was used to estimate the significance of sequence differences between *Est-6* haplotype families, treating them as geographical populations. Simulations based on the algorithms of the coalescent process with or without recombination were performed with the DnaSP and PROSEQ programs to estimate the probabilities of the observed values of Tajima's *D*, Kelly's Z_{ns} and Wall's *B* and *Q* statistics. The coalescent approach was also used to estimate confidence intervals of the nucleotide diversity values.

3. Results

3.1. Allozyme polymorphism

Electrophoretic analysis reveals two common *EST-6* allozymes: 20 lines are S, nine are F, and one exhibits a rare allozyme with slower mobility than S, denoted as US (US-255F; see Table 1). The frequencies observed (0.708 S + US and 0.292 F) are close to those from previous collections at the same site (Smit-McBride et al., 1988).

3.2. Nucleotide polymorphism and recombination

The 3062 bp sequenced comprise the 5'-flanking region (1183 bp), the complete *Est-6* coding region (1387 bp exon I, 51 bp intron, 248 bp exon II), and 193 bp of the intergenic region between *Est-6* and ψ *Est-6*. Table 1 shows a total of 79 polymorphic sites in this sample of 30 *Est-6* sequences: 28 sites in the promoter region (two associated with indels), 37

sites in the coding region (including one site in the intron), and 14 sites in the intergenic region. Four length polymorphisms (indels of 1, 5, 19 and 35 bp) are in the promoter region (Table 1). Two indels, 5-bp deletion and 35-bp insertion, have been previously observed by Odgers et al. (1995). The single nucleotide deletion (located at 383) occurs in the S-255S and S-377F strains; a 19-bp deletion (located at 348–366) occurs in the F-531F and F-611F strains.

There are 13 replacement and 23 synonymous polymorphic sites in the coding region. We have found a new replacement substitution in the US-255F (G → C, at 2372) which leads to a charge-altered amino acid replacement (amino acid position 397, neutral glycine replaced by positively charged arginine). All other detected replacement substitutions have been observed previously (Cooke and Oakeshott, 1989; Hasson and Eanes, 1996).

Table 2 shows estimates of nucleotide diversity for the entire data set, and for different haplotype families separately. In the pooled sample, total nucleotide diversity is very similar in the promoter and coding regions, but higher in the intergenic region. The level of silent variation in the coding region is higher than in the promoter region, but similar to the variation in the intergenic region, which could indicate different degrees of selective constraint in the 5'- and 3'-flanking regions. Polymorphism is lower in S than in F haplotypes (coding region) and lower in *RsaI* – than in *RsaI* + (promoter region). The difference is significant by coalescent simulations for the coding region haplotypes ($P < 0.05$), but not for the promoter haplotypes. The level of divergence (*K*) is similar in different haplotype groups within the same functional region (Table 2).

The method of Hudson and Kaplan (1985) reveals a minimum of nine recombination events in the *Est-6* region: in the intervals 433–494 (promoter); 690–1254 (promoter – exon I); 1254–1501, 1985–2032, 2032–2089, and 2093–2408 (exon I); 2408–2588 (exon I – intron); 2629–2756 (exon II); and 2803–2874 (exon II – 3'-flanking region). The *C* estimator of recombination is 0.0023 for the entire sequenced region and 0.0075 for the combined sample of all published *Est-6* sequences, including our data set (coding region) (Cooke and Oakeshott, 1989; Hasson and Eanes, 1996). *Est-6* is located within subsection 68F7–F8, on the left arm of polytene chromosome 3, a region with high recombination. The crossover rate obtained from laboratory measurements of the exchange between flanking markers is 3.32×10^{-8} /nucleotide site/generation, with maximum levels around $5\text{--}6 \times 10^{-8}$ (Comeron et al., 1999; Josep M. Comeron, personal communication). Assuming no recombination in *D. melanogaster* males and an effective population size of 10^6 , the recombination parameter is 0.0664. We note that no inversion polymorphisms have been found in the third chromosome of the El Rio population, in spite of tens of thousands of individuals sampled from different collections, including the 1991 collection used in the present study (Smit-McBride et al., 1988; Tyler et al., 1993; and our unpublished data).

Table 2

Nucleotide diversity and divergence of *Est-6* and flanking regions in 30 strains of *D. melanogaster*^a

	Promoter region	Coding region			Intergenic region	Full sequence	Silent sites
		Syn	Nsyn	Total			
Number of sites	1183	379	1256	1635	193	3062	1746
<i>Full sample:</i>							
π	0.0059	0.0147	0.0025	0.0054	0.0133	0.0062	0.0088
θ	0.0058	0.0153	0.0026	0.0056	0.0183	0.0065	0.0093
<i>K</i>	0.0833	0.1499	0.0214	0.0492	0.0955	0.0643	0.0995
<i>Slow haplotypes:</i>							
π	0.0057	0.0053	0.0008	0.0019	0.0014	0.0034	0.0053
θ	0.0054	0.0066	0.0013	0.0026	0.0029	0.0037	0.0054
<i>K</i>	0.0828	0.1504	0.0208	0.0489	0.0933	0.0639	0.0989
<i>Fast haplotypes:</i>							
π	0.0063	0.0199	0.0030	0.0069	0.0253	0.0079	0.0115
θ	0.0056	0.0194	0.0026	0.0065	0.0248	0.0074	0.0107
<i>K</i>	0.0831	0.1487	0.0227	0.0499	0.1007	0.0650	0.0997
<i>RsaI – haplotypes:</i>							
π	0.0012	0.0137	0.0026	0.0052	0.0142	0.0043	0.0055
θ	0.0025	0.0134	0.0022	0.0048	0.0117	0.0044	0.0060
<i>K</i>	0.0832	0.1499	0.0215	0.0494	0.0967	0.0645	0.0997
<i>RsaI ± haplotypes:</i>							
π	0.0046	0.0152	0.0023	0.0053	0.0104	0.0054	0.0077
θ	0.0044	0.0177	0.0025	0.0061	0.0183	0.0062	0.0089
<i>K</i>	0.0832	0.1499	0.0210	0.0490	0.0930	0.0638	0.0990

^a π is the average number of nucleotide differences per site among all pairs of sequences. θ 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. *K* is the average proportion of nucleotide differences between *D. melanogaster* and *D. simulans*, corrected according to Jukes and Cantor method. Syn, synonymous; nsyn, non-synonymous; silent, silent in non-coding regions and synonymous in coding region. The coding region includes exons I and II. Nucleotide variability is calculated for the whole data set ('full sample'), as well as for the coding (S and F haplotypes) and promoter (*RsaI* – and *RsaI* + haplotypes) regions separately. The segregating sites associated with indels are excluded from the π , θ , and *K* calculations.

3.3. Haplotype structure

There are two promoter region haplotypes, denoted as *RsaI* + and *RsaI* –, distinguished a T → G transversion at position 653 (Table 1). The average number of nucleotide differences (*K*) between the two haplotypes is 6.720. The *RsaI* – haplotypes are fairly homogeneous (*K* = 1.584), while the *RsaI* + haplotypes are more heterogeneous (*K* = 4.756). The *RsaI* – haplotypes are the most frequent in our data set (20 out of 30) and also in the data of Game and Oakeshott (1990) (20 out of 29). Odgers et al. (1995) consider the *RsaI* + to be the ancestral state, which is consistent with the higher polymorphism of the *RsaI* + haplotypes and supported by the comparison with *D. simulans* (Table 1). Three putative gene conversion events between the *RsaI* – and *RsaI* + haplotypes are revealed with the method of Betrán et al. (1997) (see Table 1).

There are also two coding region haplotypes (Table 1), denoted as the S and F allelic lineages (Balakirev et al., 1999). The average number of nucleotide differences between them is 11.809. The S group includes most haplotypes (21 out of 30), which are fairly homogeneous

(*K* = 3.810), whereas the nine F haplotypes are significantly more heterogeneous (*K* = 16.722). This high level of polymorphism would suggest that the F lineage may be ancestral, a conclusion also reached by Cooke and Oakeshott (1989) and Hasson and Eanes (1996). However, *D. simulans* has an A at position 1985, the same as the S lineage, which would support the inference that the S lineage may have been the ancestral condition from which the F allelic lineage derived, a hypothesis favored by Balakirev et al. (1999), who had in their sample only two F alleles, quite similar to one another. Four putative gene conversion events between the S and F haplotypes are revealed with the method of Betrán et al. (1997) (see Table 1). The F-775F haplotype has 12 differences with all other sequences (Table 1); an identical haplotype appears in Cooke and Oakeshott (1989; haplotype number 5), in a sample from Coff's Harbour, Australia.

The permutation test is highly significant for the promoter haplotypes $K_{st}^* = 0.4175$ ($K_{st}^*_{0.999} = 0.0966$, $P < 0.001$) as well as for the coding region haplotypes $K_{st}^* = 0.2846$ ($K_{st}^*_{0.999} = 0.0701$, $P < 0.001$). Two F haplotypes (strains F-531F and F-611F) may have arisen by recombination

Fig. 1. Distribution of nucleotide diversity within (π_{RsaI+} , π_{RsaI-} , π_F , and π_S) and between ($D_{RsaI+/RsaI-}$ and $D_{F/S}$) haplotype families. The *Est-6* sequences are divided into *RsaI* +/*RsaI* - (A); or F/S (B) groups. A schematic representation of the *Est-6* gene is displayed at bottom. Exons are indicated by black boxes; the intron and the 5'- and 3'-flanking regions are shown by thin lines. Vertical arrows indicate the locations of the *RsaI* +/*RsaI* - and F/S polymorphisms. Window size is 200 nucleotides with 50-nucleotide increments.

between S and F coding region sequences; the difference between them and the other F haplotypes is significant ($K_{st}^* = 0.3352$, $K_{st}^*_{0.95} = 0.1290$, $P < 0.05$).

3.4. Sliding window analysis

Figs. 1A,B shows the distribution of polymorphism and divergence within and between promoter haplotypes *RsaI* - and *RsaI* + and S/F alleles, respectively. Both display noticeable peaks of variation that may reflect the effect of balancing selection (Hudson and Kaplan, 1988). Fig. 1A (see Table 1) displays a distinct peak in the promoter region at coordinates 329–690, mostly due to differences between the *RsaI* - and *RsaI* + haplotypes rather than to variability within each type. The peak contains the *RsaI* site (position 653). Elevated nucleotide variation in the same promoter area of *Est-6* was revealed by Odgers et al. (1995). Fig. 1B shows two noticeable peaks, one in the promoter and one in the coding region, but only one of them (sites 1955–2093) is mostly due to differences between the S and F haplotypes. This peak includes the F/S replacement site (position 1955). We detected this peak previously (Balakirev et al., 1999) in a set of fewer sequences obtained by us, and also in the *Est-6* data of Hasson and Eanes (1996) and Cooke and Oakeshott (1989).

Fig. 2 shows a sliding window plot of the distribution of nucleotide polymorphism in *D. melanogaster* and diver-

gence between this species and *D. simulans* along the promoter (A) and coding (B) regions. The promoter peak of high polymorphism does not correspond to a region of high interspecific divergence (Fig. 2A), which is inconsistent with the neutral model. The peak contains a stretch of polymorphic sites at positions 494, 518, and 544 (Table 1) that are in significant linkage disequilibrium with the *RsaI* site (position 653). In the coding region (Fig. 2B) there is also discordance between polymorphism and interspecific divergence, within a segment that includes the polymorphic sites at positions 1955, 1985, 2032, 2089, 2092, and 2093 (Table 1), which are in significant linkage disequilibrium (position 1955 determines, as noted, the F/S polymorphism).

We measure heterogeneity in the distribution of silent polymorphic sites relative to fixed interspecific differences and by means of Goss and Lewontin (1996) and McDonald (1998) statistics and assess their significance by Monte Carlo simulations of the coalescent model incorporating recombination (McDonald, 1998). We use models with recombination because intragenic recombination predicts heterogeneity in the level of diversity along the sequence. Based on 10,000 simulations, with the recombination parameters varying from 1 to 64, Goss and Lewontin (1996) statistics are significant for the whole *Est-6* gene (3062 bp): interval length variance (V_{IL}) is 0.000353, $P = 0.031$

Fig. 2. Sliding window plots of nucleotide diversity (π ; continuous line) and nucleotide divergence (K ; dotted line), across the *Est-6* 5'-flanking (A); and coding (B) regions. Nucleotide divergence is estimated as the average number of nucleotide substitutions per site between *D. melanogaster* and *D. simulans*. The K values are divided by $T + 1 = 9.885$, where T is the divergence time between *D. melanogaster* and *D. simulans*, measured in $2N$ generations (N is the effective population size (Hudson et al., 1987). Window sizes are 80 nucleotides with 20-nucleotide increments for A, and 130 nucleotides with 25-nucleotide increments for B.

and modified interval length variance (Q_{IL}) is 0.000947, $P = 0.047$. For the promoter region alone the tests do not reveal significant heterogeneity in the polymorphism-to-divergence ratio. However, for our data set combined with previously published *Est-6* promoter sequences (Odgers et al., 1995), McDonald (1998) maximum sliding G statistic (G_{max}) as well as Goss and Lewontin (1996) statistics are significant: $G_{max} = 18.0264$, $P = 0.025$; $V_{IL} = 0.00232$, $P = 0.013$, $Q_{IL} = 0.00471$, $P = 0.030$. The tests fail to reveal significant heterogeneity in the polymorphism-to-divergence ratio for the *Est-6* coding region.

3.5. Linkage disequilibrium

Fig. 3 shows the significance levels of Fisher exact test for

linkage disequilibrium between non-singleton pairs of segregating sites. Out of 1225 pairwise comparisons, 445 (36.3%) show significant disequilibrium. With the Bonferroni correction for multiple comparisons, there remain 114 (9.3%) significant associations. There are two areas with significant associations between sites: one includes the promoter region; the other includes the rest of the gene. The significant associations within each region are 52.9 and 55.2%, respectively (or 33.3 and 11.5% with the Bonferroni correction). The significant associations between the two regions are substantially smaller, 15.1% (or 1.0% with the Bonferroni correction), and they are mostly due to the site at position 527. Site 527 is anomalous in that it is in linkage equilibrium with all other sites in the promoter region, but is in disequilibrium with sites in the rest of the gene.

Fig. 3. Fisher exact test of non-random associations between pairs of *Est-6* polymorphisms. Singleton mutations are excluded from the analysis. Each box in the matrix represents the comparison of two polymorphic sites. The location of the segregating sites on the *Est-6* gene is shown on the diagonal, which indicates the position of the 5'-flanking, coding, and 3'-flanking regions.

Table 3
McDonald-Kreitman test of neutrality for the *Est-6* gene

	Fixed differences	Polymorphisms		
		<i>D. melanogaster</i>	<i>D. simulans</i> ^a	Total ^b
Replacement	18	13	11	22
Silent	24	23	54	74

^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 $P = 0.025$.

3.6. Tests of neutrality

In our previous study we detected significant deviations from neutrality in the coding region of *Est-6* using the McDonald and Kreitman (1991) test applied to four *D. simulans* (Karotam et al., 1995) and 15 *D. melanogaster* sequences (Balakirev et al., 1999). Similar results are obtained for the 30 *D. melanogaster Est-6* coding-region sequences (Table 3). However the McDonald and Kreitman (1991) test is not significant if we combine our data with published *Est-6* coding sequences (Cooke and Oakeshott, 1989; Hasson and Eanes, 1996) (Fisher's exact test $P = 0.074$). Hudson et al. (1987); Tajima (1989); Fu and Li (1993), and Depaulis and Veuille (1998) tests do not reveal any significant deviation from neutrality for the *Est-6* region. However, 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 a significant deviation from neutrality expectations for the S alleles ($D = -1.864$, $P < 0.05$), but not for the F alleles ($D = -0.181$, $P > 0.10$), which is consistent with our previous results (Balakirev et al., 1999). In the promoter region we detect a significant deviation from neutrality for the *RsaI* – haplotypes using both Tajima ($D = -2.065$, $P < 0.05$) and Fu and Li ($D^* = -2.672$, $P < 0.05$; $F^* = -2.890$, $P < 0.05$) tests. Negative values of the Tajima's D statistic indicate an excess of polymorphisms segregating at low frequency, and negative values of the Fu and Li's statistics indicate an excess of unique polymorphisms in the data set. We have also applied Kelly's Z_{ns} (1997) and Wall (1999) B and Q tests, which are based on linkage disequilibrium between segregating sites. For the entire *Est-6* region both tests are highly significant ($Z_{ns} = 0.154$, $P = 0.004$; $B = 0.270$, $P = 0$; $Q = 0.427$, $P = 0$) with $C = 0.015$, equal to the C_{min} based on the inferred minimum number of the recombination events (Hudson and Kaplan, 1985). The tests are also significant for the coding and promoter regions separately, with $C \geq 0.020$ (promoter region) and $C \geq 0.010$ (coding region) for Kelly's test and $C \geq 0.010$ for Wall's tests. The areas of significant values of Kelly's and Wall's statistics coincide with the peaks of linkage disequilibrium (data not shown) and are centered around the *RsaI* site (position 653, promoter region), three replace-

ment sites in the beginning of exon I (positions 1212, 1254, and 1356), and around the F/S polymorphism (position 1955) (Table 1, Fig. 3). Thus the neutrality tests incorporating recombination (Kelly, 1997; Wall, 1999) are significant for the *Est-6* region, with recombination rate below the laboratory estimate $C_{lab} = 0.0664$.

3.7. Divergence time

The *Est-6* average distances (nucleotide differences) between *D. simulans* and *D. melanogaster* for the promoter and coding plus 3'-flanking region are 75.9 and 97.7, respectively. If we assume that the species diverged 2.3–2.5 million years ago, the time of divergence for the *RsaI* – promoter haplotypes and S coding-region haplotypes, under the assumption of the molecular clock, are about 50 and 94 thousand years, respectively. We have also calculated the expected time of divergence within and between the haplotype families for both regions following the method of Hudson et al. (1997), which infers the time of a putative selective sweep from the amount of variation that has accumulated within the relatively homogeneous subset of sequences. For the promoter region the expected divergence time for the *RsaI* – haplotypes is about 25 thousand years (excluding the three recombinant sites at positions 338, 368, and 433). For the coding region the expected divergence time within the S haplotypes is about 64 thousand years (without the four recombinant sites at positions 1254, 1985, 2032, and 2408). The time of divergence between the two main promoter allelic lineages (*RsaI* + and *RsaI* –) is 50 thousand years with Hudson et al. (1997) method, but more than seven times greater, 380 thousand years, under the molecular clock assumption. These discrepancies can be accounted for by natural selection that accelerates the evolution of the selectively favored alleles. Parallel differences occur for the coding region, but the differences between the estimates obtained by the two methods are smaller. Thus, the estimated age of the *Est-6* S allelic lineage is 64 thousand years or 94 thousand years, and the divergence between the S-F lineages is 162 thousand or 485 thousand years, depending on the method used. This is consistent with the hypothesis that the divergence between the two coding-region allelic lineages has been impelled by natural selection, but with lesser strength than in the case of the promoter allelic lineages.

4. Discussion

4.1. Pattern of variability

The pattern of *Est-6* nucleotide variability is highly structured and has distinctive features in the promoter and coding regions (Table 1). The promoter pattern of variation does not carry over into the coding region, but the pattern of variability in the 3'-flanking region is similar to the coding region. The level of silent diversity is similar in the coding

and 3'-flanking regions, but substantially lower in the promoter region (Table 2), which may reflect stronger purifying selection for the promoter region. Strong linkage disequilibrium occurs within each of the regions, but is less pronounced between them (Fig. 3).

We previously suggested that the pattern of nucleotide variability of the *Est-6* coding region is shaped by the influence of both directional and balancing selection (Balakirev et al., 1999). The present analysis of the *Est-6* coding region confirms our previous hypothesis. The pattern of variability in the promoter region of the gene also suggests involvement of both forms of selection (see Section 3). It has been repeatedly shown that positive selection is involved in the evolution of sex-related genes in *Drosophila*, as well as other organisms (reviewed by Civetta and Singh, 1999). A situation in which both directional and balancing selection might be involved in shaping nucleotide variability has been proposed for the *Acp29AB*, also a sex-related gene of *D. melanogaster* (Aguadé, 1999).

4.2. Target of selection

Game and Oakeshott (1990) revealed that the *RsaI* promoter polymorphism significantly influences EST-6 activity: males of lines lacking the *RsaI* site show 25% lower EST-6 activity than lines with the site. This reduction impacts the reproductive fitness of *D. melanogaster* (see Oakeshott et al., 1995). The pattern of association between EST-6 activity and reproductive fitness is complex, but leaves no doubt about the adaptive nature of the EST-6 activity variation. We have observed a pronounced area of highly significant linkage disequilibrium around the *RsaI* site (Fig. 3), which is manifest also by Kelly's and Wall's statistics (data not shown). Thus the *RsaI* site (position 653, Table 1) might be a target of selection in the promoter region. However this site is in linkage disequilibrium with 11 other polymorphic sites (Fig. 3), which makes it impossible to ascertain the precise site that impacts EST-6 activity. Conceivably, selection could act on any site that is in linkage disequilibrium with the *RsaI* site, or on the whole stretch of linked sites. Seeking to localize more precisely the putative selection site, we have examined the pattern of polymorphism and divergence in the promoter region using a sliding-window. There is a region with apparent excess of polymorphism, which is associated with a remarkable low level of interspecific divergence (Fig. 2A). This region includes the stretch of the polymorphic sites at positions 494, 518, and 544 (Table 1), which are in linkage disequilibrium with the *RsaI* site (position 653) (Fig. 3).

A possible interpretation is that directional selection affects the *RsaI* - haplotypes, but without complete exclusion of the *RsaI* + haplotypes. This can be explained by the fact that the *RsaI* + mutation is associated with a significant increase of EST-6 activity, which has important consequences for reproductive fitness (Oakeshott et al., 1995). From the analysis of the correlations between EST-6 activ-

ity and reproductive fitness it can be expected that both *RsaI* + and *RsaI* - strains may have some advantages under particular conditions (Oakeshott et al., 1995). Thus the *RsaI* - / *RsaI* + polymorphism may be a balanced polymorphic system, which is indirectly confirmed by the peak of nucleotide variability in the corresponding promoter region detected in our data (Fig. 1A), as well as in the data of Odgers et al. (1995).

There are some lines of evidence indicating that the *Est-6* coding-region variation is also subject to positive selection (Oakeshott et al., 1995; Richmond et al., 1990; Balakirev et al., 1999). Cooke and Oakeshott (1989) propose that one or both of the polymorphisms at positions 1955 or 1985 (our coordinates) are the primary targets for the selection underlying the F/S allozyme latitudinal clines. Analysis of DNA sequence variation in *Est-6* reveals an excess of polymorphism surrounding the two replacement substitutions (Balakirev et al., 1999; Fig. 1B). The area of noticeable differences between observed and expected levels of variation (Fig. 2B) includes site 1955, which determines the F/S allozyme polymorphism (Table 1). There is an area with elevated level of linkage disequilibrium in the coding region (Fig. 3), accompanied by highly significant values of Kelly's and Wall's statistics (data not shown). The evidence also suggests that the S allelic group might be subject to directional selection (see Section 3; see also Cooke and Oakeshott, 1989; Oakeshott et al., 1995; Balakirev et al., 1999).

4.3. Evolutionary scenario

A possible interpretation starts with a favorable mutation in the promoter region spread by directional selection; a second promoter haplotype would persist in populations because of balancing selection involving both haplotypes. Similar processes may have occurred in the coding region. If both favorable mutations are present in the same DNA segment including the promoter and coding region, then the corresponding haplotype would be under a double selective sweep, with greater effect in increasing its frequency ('double sweep' haplotypes). In our data, strains S-5F, S-26F, S-94F, S-174F, US-255F, S-483F, S-498F, S-521F, S-565F, S-581F, S-968F, S-1224F, S-438S, and S-521S have the mutations that are putatively under selection in both the promoter and the coding region. The 'double sweep' haplotypes are frequent (14 out of 30) and have minimum variability along the whole length of the DNA segment studied. The average number of nucleotide differences (*K*) between them is low (4.47). Tajima's *D* is negative and significant for this group of haplotypes: *D* = -1.809, *P* < 0.05. The haplotype test (Hudson et al., 1994) is also significant (*P* < 0.05) for the 'double sweep' haplotypes including singleton polymorphisms with recombination *C* ≥ 0.04. The haplotypes that have just one mutation under directional selection ('single sweep' haplotypes) located either in the promoter (F-517S, F-1461S, F-274F, F-357F, F-517F,

F-611F) or in the coding (S-114S, S-255S, S-501S, S-510S, S-549S, S-2588S, S-377F) region have lower frequency, higher variability (K is equal to 10.80 and 10.19 for the promoter and coding region ‘single sweep’ haplotypes, respectively), and less pronounced indication of a selective sweep. The haplotypes that have no mutation subject to putative selection (‘zero sweep’ haplotypes) are most variable ($K = 28.00$) and less frequent (three out of 30: F-96S, F-531F, F-775F), which may reflect their old age.

According to this scenario, the pattern of nucleotide variation in *Est-6* is shaped by the superposition of the effects of directional and balancing selection in the promoter region and by analogous superposition effects in the coding region. The differential hitchhiking of the mutations incorporated in the different haplotypes of the two regions complicates the observed pattern of polymorphism. It is, nevertheless, possible to recognize some general pattern in the network constructed for the combined promoter and coding region (exon I) (Fig. 4). The left part of the network includes the S haplotypes, whereas the right part contains the F haplotypes. The *RsaI* + and *RsaI* – haplotypes

occupy the upper and lower parts of the network, respectively (divided by the horizontal dotted line). The ‘double sweep’ haplotypes are in the *RsaI* – /S compartment (lower left). The ‘single sweep’ haplotypes are in the *RsaI* + /S (coding region sweep; upper left) and *RsaI* – /F (promoter region sweep; lower right) compartments. The ‘zero sweep’ haplotypes are in the *RsaI* + /F compartment (upper right).

Note that the concentration of F haplotypes in the *RsaI* – /F compartment is maintained by selection in the promoter region whereas the concentration of the *RsaI* + haplotypes (*RsaI* + /S compartment) is maintained by selection in the coding region (Fig. 4, Table 1). This observation contradicts the conclusion of Odgers et al. (1995), who reject the possibility that previously reported latitudinal clines in the *Est-6* allozyme frequencies are due to their hitchhiking along with selection on the promoter difference. But allozyme latitudinal clines could be generated by the interaction of selective processes in the promoter and coding regions and by the rate of recombination between them, because the S haplotypes are under a double sweep (*RsaI* – /S compartment) and a single sweep (*RsaI* + /S

Fig. 4. One-step network of the *Est-6* sequences including the 5'-flanking region and exon I. To generate the network the sequence data are preprocessed into binary characters. The haplotypes included within circles are observed haplotypes; small solid circles represent hypothetical intermediate haplotypes. Identical haplotypes are enclosed together. Solid connections between haplotypes indicate transitions that are unambiguous, whereas the thin dashed lines indicate hypothetical connections. Numbers along the lines represent mutational steps. There are four compartments, designated as *RsaI* + /S, *RsaI* + /F, *RsaI* – /S, and *RsaI* – /F, divided by the horizontal and vertical dotted lines.

compartment) influence, whereas the F haplotypes are only under a single sweep (*RsaI* – /F compartment) influence.

There are, nevertheless, non-selective hypotheses that could account for the patterns of *Est-6* polymorphism. In particular, population dynamics involving bottlenecks and founding effects and/or population admixture could be the explanatory processes. One way to distinguish between the selective and the demographic hypotheses may be to compare our results with those obtained with other locus (or loci) investigated in the same lines of *D. melanogaster* as well as in other populations. Whether or not such information will change the selective hypotheses favored by the current data remains to be seen.

Acknowledgements

We are grateful to D.A. Filatov, J.K. Kelly, J.H. McDonald, J.D. Wall, J.M. Comeron, F. Depaulis, and J. Rozas for useful advice on analyzes and for providing computer programs. J.M. Comeron generously provided the estimates of the recombination rates. We thank Andrei Tatarenkov, Victor DeFilippis, Martina Zurovkova, and Carlos Márquez for comments and help; and Narelle Dryden, W.M. Fitch, B. Gaut, R.R. Hudson, A. Long, and one anonymous reviewer for detailed and valuable comments. This work is supported by NIH grant GM42397 to F.J. Ayala.

References

- Aguadé, M., 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* 152, 543–551.
- Balakirev, E.S., Ayala, F.J., 1996. Is esterase-P encoded by a cryptic pseudogene in *Drosophila melanogaster*? *Genetics* 144, 1511–1518.
- Balakirev, E.S., Balakirev, E.I., Rodríguez-Trelles, F., Ayala, F.J., 1999. Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. *Genetics* 153, 1357–1369.
- Betrán, E., Rozas, J., Navarro, A., Barbadilla, A., 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* 146, 89–99.
- Civetta, A., Singh, R.S., 1999. Broad-sense sexual selection, sex gene pool evolution, and speciation. *Genome* 42, 1033–1041.
- Collet, C., Nielsen, K.M., Russell, R.J., Karl, M., Oakeshott, J.G., Richmond, R.C., 1990. Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. *Mol. Biol. Evol.* 7, 9–28.
- Comeron, J.M., Kreitman, M., Aguadé, M., 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* 151, 239–249.
- Cooke, P.H., Oakeshott, J.G., 1989. Amino acid polymorphisms for esterase-6 in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 86, 1426–1430.
- Depaulis, F., Veuille, M., 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.* 15, 1788–1790.
- Filatov, D.A., Charlesworth, D., 1999. DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. *Genetics* 153, 1423–1434.
- Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.
- Game, A.Y., Oakeshott, J.G., 1990. Associations between restriction site polymorphism and enzyme activity variation for esterase 6 in *Drosophila melanogaster*. *Genetics* 126, 1021–1031.
- Goss, P.J.E., Lewontin, R.C., 1996. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* 143, 589–602.
- Hasson, E., Eanes, W.F., 1996. Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. *Genetics* 144, 1565–1575.
- Hudson, R.R., Kaplan, N., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164.
- Hudson, R.R., Kaplan, N., 1988. The coalescent process in models with selection and recombination. *Genetics* 120, 831–840.
- Hudson, R.R., Kreitman, M., Aguadé, M., 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159.
- Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowski, J., Ayala, F.J., 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* 136, 1329–1340.
- Hudson, R.R., Sáez, A.G., Ayala, F.J., 1997. DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. *Proc. Natl. Acad. Sci.* 94, 7725–7729.
- Karotam, J., Boyce, T.M., Oakeshott, J.G., 1995. Nucleotide variation at the hypervariable esterase 6 isozyme locus of *Drosophila simulans*. *Mol. Biol. Evol.* 12, 113–122.
- Kelly, J.K., 1997. A test of neutrality based on interlocus associations. *Genetics* 146, 1197–1206.
- McDonald, J.H., 1998. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol. Biol. Evol.* 15, 377–384.
- McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654.
- Oakeshott, J.G., Boyce, T.M., Russell, R.J., Healy, M.J., 1995. Molecular insights into the evolution of an enzyme – esterase 6 in *Drosophila*. *TREE* 10, 103–110.
- Odgers, W.A., Healy, M.J., Oakeshott, J.G., 1995. Nucleotide polymorphism in the 5' promoter region of esterase 6 in *Drosophila melanogaster* and its relationship to enzyme activity variation. *Genetics* 141, 215–222.
- Richmond, R.C., Nielsen, K.M., Brady, J.P., Snella, E.M., 1990. Physiology, biochemistry and molecular biology of the *Est-6* locus in *Drosophila melanogaster*. In: Barker, J.S.F., Starmer, W.T., MacIntyre, R.J. (Eds.). *Ecological and Evolutionary Genetics of Drosophila*, Plenum Press, New York, pp. 273–292.
- Rozas, J., Rozas, R., 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 15, 174–175.
- Seager, R.D., Ayala, F.J., 1982. Chromosome interactions in *Drosophila melanogaster*. I. Viability studies. *Genetics* 102, 467–483.
- Smit-McBride, Z., Moya, A., Ayala, F.J., 1988. Linkage disequilibrium in natural and experimental populations of *Drosophila melanogaster*. *Genetics* 120, 1043–1051.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- Tamarina, N.A., Ludwig, M.Z., Richmond, R.C., 1997. Divergent and conserved features in the spatial expression of the *Drosophila pseudoobscura esterase-5B* gene and the *esterase-6* gene of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 94, 7735–7741.
- Tyler, R.H., Brar, H., Singh, M., Latorre, A., Graves, J.L., Mueller, L.D., Rose, M.R., Ayala, F.J., 1993. The effect of superoxide dismutase alleles on aging in *Drosophila*. *Genetica* 91, 143–149.
- Wall, J.D., 1999. Recombination and the power of statistical tests of neutrality. *Genet. Res.* 74, 65–79.