## Title

DNA polymorphism in the beta-Esterase gene cluster of Drosophila melanogaster

Journal
Genetics, 164(2)

ISSN
0016-6731

Authors
Balakirev, E S
Chechetkin, V R
Lobzin, V V
et al.

Publication Date
2003-06-01

Peer reviewed

# DNA Polymorphism in the β-*Esterase* Gene Cluster of *Drosophila melanogaster*

**Evgeniy S. Balakirev,[*,†] V. R. Chechetkin,[‡] V. V. Lobzin[‡] and Francisco J. Ayala[*,1]**

*[*]Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, [†]Institute of Marine Biology, Vladivostok 690041, Russia, [†]Academy of Ecology, Marine Biology, and Biotechnology, Far Eastern State University, Vladivostok 690600, Russia and [‡]Troitsk Institute of Innovation and Thermonuclear Investigations (TRINITI), Theoretical Department of Division for Perspective Investigations, 142190 Troitsk, Moscow Region, Russia*

Manuscript received December 16, 2002
Accepted for publication February 21, 2003

## ABSTRACT

We have analyzed nucleotide polymorphism within a 5.3-kb region encompassing the functional *Est-6* gene and the ψ*Est-6* putative pseudogene in 28 strains of *Drosophila melanogaster* and one of *D. simulans*. Two divergent sequence types were detected, which are not perfectly associated with *Est-6* allozyme variation. The level of variation (π) is very close in the 5′-flanking region (0.0059) and *Est-6* gene (0.0057), but significantly higher in the intergenic region (0.0141) and putative pseudogene (0.0122). The variation in the 3′-flanking region is intermediate (0.0083). These observations may reflect different levels of purifying selection in the different regions. Strong linkage disequilibrium occurs within the region studied, with the largest values revealed in the putative pseudogene and 3′-flanking region. Moreover, recombination is restricted within ψ*Est-6*. Gene conversion is detected both within and (to a lesser extent) between *Est-6* and ψ*Est-6*. The data indicate that ψ*Est-6* exhibits some characteristics that are typical of nonfunctional genes, while other characteristics are typically attributed to functional genes; the same situation has been observed in other pseudogenes (including Drosophila). The results of structural entropy analysis demonstrate higher structural ordering in *Est-6* than in ψ*Est*-6, in accordance with expectations if ψ*Est*-6 is indeed a pseudogene. Taking into account that the function of ψ*Est-6* is not known (but could exist) and following the terminology of J. Brosius and S. J. Gould, we suggest that the term "*poto*gene" may be appropriate for ψ*Est-6*, indicating that it is a *pot*ential gene that may have acquired some distinctive but unknown function.

THE β-*esterase* gene cluster is on the left arm of chromosome 3 of *Drosophila melanogaster*, at 68F7–69A1 in the cytogenetic map (but see Procunier *et al.* 1991). The cluster is composed of two tandemly duplicated genes, originally named *Est-6* and *Est-P* (Collet *et al.* 1990). The coding regions of these genes are 1686 and 1691 bp long, respectively, and consist of two exons (1387 and 248 bp) and a small (51 bp in *Est-6* and 56 bp in *Est-P*) intron (Oakeshott *et al.* 1987). The *Est-6* gene is well characterized (reviewed by Richmond *et al.* 1990; Oakeshott *et al.* 1993). The gene encodes the major β-carboxylesterase (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). Less information is available for *Est-P*. Collet *et al.* (1990) first described *Est-P* and concluded that it was a functional gene, on the basis of several lines of evidence: transcriptional activity, intact splicing sites, no premature termination codons, and presence of initiation and termination codons. However, Balakirev and Ayala (1996) found premature stop codons within the *Est-P* coding region and some other indications suggesting that *Est-P* might be in fact a pseudogene and named it ψ*Est-6*. Dumancic *et al.* (1997) showed that some alleles of the *Est-P* produce a catalytically active esterase corresponding to the previously identified EST-7 isozyme (Healy *et al.* 1991) and renamed the gene correspondingly *Est-7*.

Our earlier investigation of ψ*Est-6* (Balakirev and Ayala 1996) was limited to 10 lines of *D. melanogaster* and 2.8 kb. We now increase the sample size to 28 lines and extend the analysis by comparing the nucleotide variability in the ψ*Est-6* putative pseudogene and *Est-6* gene in a random sample of *D. melanogaster* derived from a natural population of California. The full sequence now analyzed is 5394 bp long and includes the 5′-flanking region, complete *Est-6* gene, intergenic region, ψ*Est-6* putative pseudogene, and 3′-flanking region. The data for the 5′-flanking region and *Est-6* gene (1686 bp) are from Balakirev *et al.* (2002).

## MATERIALS AND METHODS

**Drosophila strains:** The 28 *D. melanogaster* strains were derived from a random sample of wild flies collected by F. J. Ayala (October 1991) in El Rio Vineyard, Acampo, California. 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 esterase-6 (the letter before the hyphen) and superoxide

[1] *Corresponding author:* Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525. E-mail: fjayala@uci.edu

dismutase (the letter after the number) electrophoretic alleles they carry, ultra slow (US), slow (S), and fast (F) (Figure 1).

**DNA extraction, amplification, and sequencing:** Total genomic DNA was extracted using the tissue protocol of the QIAamp tissue kit (QIAGEN, Valencia, CA). The *D. melanogaster Est-6* sequence (GenBank accession nos. M33780 and M33781; Collet *et al.* 1990) was used for designing PCR and sequencing primers. The primers used for the PCR amplification reactions were 5′-gattttgcttcgagtgataatgg-3′ (forward primer) and 5′-agactacgtgcacagtgtggtggg-3′ (reverse primer). The PCR reactions were carried out in final volumes of 50 μl using TaKaRa Ex Taq in accordance with the manufacturer's description (Takara Biotechnology, Berkeley, CA). The reaction mixtures were subjected to 30 cycles of denaturation, annealing, and extension: 95° for 30 sec, 63° for 30 sec, and 72° for 2.0 min (for the first cycle and progressively adding 3 sec at 72° for every subsequent cycle); with a final 7-min extension period at 72°. The PCR reactions were purified with the Wizard PCR preps DNA purification system (Promega, Madison, WI), directly sequenced by the dideoxy chain-termination technique using Dye Terminator chemistry, and separated with the ABI PRISM 377 automated DNA sequencer (Perkin-Elmer, Norwalk, CT). For each line, the sequences of both strands were determined, using 12 overlapping internal primers spaced, on average, 350 nucleotides. (See GenBank accession nos. AF147095–AF147102, AF150809–AF150815, AF217624–AF217645, and AF526538–AF526558 for the ψ*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.

**DNA sequence analysis:** The sequences were assembled using the program SeqMan (Lasergene, DNASTAR, 1994–1997). 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 analyses. Departures from neutral expectations were investigated using Kelly's (1997) and Wall's (1999) neutrality tests incorporating recombination. The permutation approach of Hudson *et al.* (1992) was used to estimate the significance of sequence differences between haplotype families. The coalescent simulations (Hudson 1990) were performed with the PROSEQ program to estimate the probabilities of the observed values of Kelly's $Z_{nS}$ and Wall's $B$ and $Q$ statistics and confidence intervals for the nucleotide diversity values. The method of Sawyer (1989, 1999) was used to analyze intra- and intergenic conversion events.

**Entropy analysis:** If ψ*Est-6* is in fact a pseudogene or nonessential gene, one could expect lower structural regularity and higher structural divergence in this putative pseudogene than in its functional paralogous gene, *Est-6*. These features can be quantitatively assessed with the proper structural analysis of the relevant sequences. Our approach is based on spectral methods previously developed (Chechetkin and Turygin 1994, 1996; Chechetkin *et al.* 1994; Chechetkin and Lobzin 1996, 1998; for a review and further references, see Lobzin and Chechetkin 2000).

First, we begin with the necessary definitions. The Fourier harmonics corresponding to the nucleotides of type α (where α is A, C, G, or T) in a sequence of length $M$ are defined as

$$\rho_\alpha(q_n) = M^{-1/2}\sum_{m=1}^{M}\rho_{m,\alpha}e^{-iq_n m}, \quad q_n = 2\pi n/M, \quad n = 0,1,\ldots, \quad M-1, \tag{1}$$

where $\rho_{m,\alpha}$ indicates the positions occupied by the nucleotides of type α, $\rho_{m,\alpha} = 1$ if the nucleotide of type α occupies the $m$th site, and 0 otherwise. The amplitudes of Fourier harmonics (or structure factors) are expressed as

$$F_{\alpha\alpha}(q_n) = \rho_\alpha(q_n)\rho^*_\alpha(q_n), \tag{2}$$

where the asterisk denotes complex conjugation. The zeroth harmonics depending only on the nucleotide composition do not contain structural information and are discarded below. Due to the symmetry property

$$F_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(2\pi - q_n), \tag{3}$$

the spectra for structure factors can be restricted to their left halves from $n = 1$ to

$$N = [M/2], \tag{4}$$

where the brackets denote the integer part of the quotient. The structure factors will always be normalized with respect to mean spectral values,

$$f_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(q_n)/\bar{F}_{\alpha\alpha}; \quad \bar{F}_{\alpha\alpha} = N_\alpha(M - N_\alpha)/M(M-1), \tag{5}$$

where $N_\alpha$ is the total number of nucleotides of type α in a sequence of length $M$. The structural regularity of the nucleotides of type α in a sequence of length $M$ is assessed with the spectral structural entropy

$$S_\alpha = -\sum_{n=1}^{M-1} f_{\alpha\alpha}(q_n)\ln f_{\alpha\alpha}(q_n). \tag{6}$$

The value of spectral entropy for a counterpart random sequence having the same nucleotide composition is the highest and the corresponding mean characteristics averaged over an ensemble of various random realizations are given by

$$\langle S_\alpha\rangle_{\text{random}} \cong -0.422785\ldots(M-1);$$
$$\langle(\Delta S_\alpha)^2\rangle_{\text{random}} \cong 0.579736\ldots(M-1). \tag{7}$$

Using the values (7) as the reference characteristics, it is convenient to introduce the relative spectral structural entropy

$$S_{\alpha,\text{rel}} = (\langle S_\alpha\rangle_{\text{random}} - S_\alpha)/|\langle S_\alpha\rangle_{\text{random}}| \tag{8}$$

as well as the relative normalized deviations

$$r_\alpha = (\langle S_\alpha\rangle_{\text{random}} - S_\alpha)/\langle(\Delta S_\alpha)^2\rangle^{1/2}_{\text{random}}. \tag{9}$$

The value of $S_{\alpha,\text{rel}}$ serves for the comparison of structural regularity for the different sequences (generally, also of different lengths); the higher the value of $S_{\alpha,\text{rel}}$, the higher the structural regularity of a sequence, while $r_\alpha$ serves for the assessment of the statistical significance of observed deviations. Assuming a Gaussian distribution for $r$ in the case of random deviations, the probability of finding values of $r$ exceeding some threshold $r_0$ is given by

$$\Pr(r > r_0) = \frac{1}{\sqrt{2\pi}}\int_{r_0}^{\infty} e^{-r^2/2}\,dr. \tag{10}$$

The value $\Pr = 0.05$ corresponds to that of $r_0 = 1.64$.

The level of structural divergence may be quantitatively estimated with deviations

$$\delta_\alpha = 1 - k_{\alpha\alpha}(1|2), \tag{11}$$

where the cross correlation coefficients are determined as

$$k_{\alpha\alpha}(1|2) = \frac{\sum_{n=1}^{N}(F^{(1)}_{\alpha\alpha}(q_n) - \bar{F}^{(1)}_{\alpha\alpha})(F^{(2)}_{\alpha\alpha}(q_n) - \bar{F}^{(2)}_{\alpha\alpha})}{N\sigma(F^{(1)}_{\alpha\alpha})\sigma(F^{(2)}_{\alpha\alpha})}, \tag{12}$$

$$\sigma^2(F_{\alpha\alpha}) = N^{-1}\sum_{n=1}^{N}(F_{\alpha\alpha}(q_n) - \bar{F}_{\alpha\alpha})^2. \tag{13}$$

The structure factor harmonics $F_{\alpha\alpha}(q_n)$, the mean spectral values $\bar{F}_{\alpha\alpha}$, and number $N$ are defined in Equations 2, 4, and 5, while the superscripts 1 and 2 refer to a pair of compared

```
                                                           1111111111111111111111111111111111111112222222
                11111223333344444445555556666778888999999990000122223333333445555556666667778888890000012
            168802455330247823568890014454567030127233455551569212393677789130008991378891263457941237921
           8267891858234412788397884501209394554333437201271188902989547980765671245430676337069410081638
S-255S     TCTAAATTAA†ATTAGCTAACGCCACGCTCTATCAAGGCTCAGCCAGTCTGTGGATGGCAT‡TCCCTACAACAGAGGCAGCCCTAAAGTTACCT
           **  ** *** ** *  *  ***  *  ** ***    *   * * *  ****** **    **   *** **  **
S-438S     ..............................................................................................
S-377F     ..............................................................................................
S-581F     ..............................................................................................
S-94F      ...........G..................................................................................
S-174F     .....C.....G..................................................................................
US-255F    .....C.....G..................................................................................
F-611F     .......A......................................................................................
F-96S      .......A.....................................................................................C
S-521S     .......A.....................................................................................C
S-521F     .......A.....................................................................................C
S-1224F    .......A.....................................................................................C
S-501S     ....................T........................................▼G...............................
S-510S     ....................T........................................▼G...............................
S-5F       ....................T........................................▼G...............................
F-531F     ...........G................C...........................A.....................................
S-565F     ...........GC......................................................................A.....A....
S-26F      ............C...........G..........................................................A.....A....
S-968F     ............C...........G..........................................................A.....A....
S-483F     ...TT.C..C.G.................................................................................C
S-114S     .................A................................................AGTGGGGAGA..GAAAT.GG...AGT.C
S-549S     .................A................................................AGTGGGGAGA..GAAAT.GG...AGT.C
F-357F     CAATT.C..C.....TT....................A.CTTCTAGTCT.A...GC.AGCC..AGTAGTGGGGAGA..GAAAT.GG...AGT.C
F-274F     CAATT.C..C.....TT.CG.TTTGGAT.AC.GTCT.A.CTTCTAGTCT.A...GC.AGCC..AGTAGTGGGGAGA..GAAAT.GG...AGTTC
F-517S     CAATT.C..C.....TT.CG.TTTGGAT.AC.GTCT.A.CTTCTAGTCT.A...GC.AGCC..AGTAGTGGGGAGA..GAAAT.GG...AGTTC
F-517F     CAATT.C..C.....TT.CG.TTTGGAT.AC.GTCT.A.CTTCTAGTCT.A...GC.AGCC..AGTAGTGGGGAGA..GAAAT.GG...AGTTC
F-1461S    CAATT.C..C.....TT.CG.TTTGGAT.AC.GTCT.A.CTTCTAGTCT.A...GC.AGCC..AGTAGTGGGGAGA..GAAAT.GG...AGTTC
F-775F     ..ATT.C.TC▲..CG........TGG....CGG.CTA.TC.........A.ATAGC..GCC..AGTAGTGGGGAG.TTGA......TAC..T.C
```

FIGURE 1.—DNA polymorphism in the putative pseudogene ψ*Est-6*. The numbers above the top sequence represent the segregating sites and the start of a deletion or insertion. Nucleotides are numbered from the *Est-P* start codon (position 3052 in COLLET *et al.* 1990). Eleven premature stop codons are due to single-nucleotide polymorphisms G/A (site 1068, strains F-357F, F-274F, F-517F, F-517S, and F-1461S) and T/G (site 1388, strains S-501S, S-510S, and S-5F), as well as to a 9-bp insertion of ACATTTGAT (position 1379–1387, strains S-501S, S-510S, and S-5F); these sites and nucleotides are marked by boldface type. The S, US, and F letters before the strain 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 and have been previously used to tag these lines.) ▲ denotes a 3-bp deletion of CAG (position 232–234, strain F-775F); † denotes the absence of a deletion; ▼ denotes an insertion of ACATTTGAT (position 1379–1387, strains S-501S, S-510S, and S-5F); ‡ denotes the absence of an insertion.

sequences. The definition (12) assumes equal lengths for the compared sequences 1 and 2 corresponding to the patterns of the same gene in two different strains of *D. melanogaster*. In the presence of insertions/deletions and unequal lengths in the compared sequences, the shorter sequence is supplemented by void sites up to the length of the longer one (MARPLE 1987). For $n$ different sequences there are $n(n - 1)/2$ pairwise cross correlation coefficients. In our case $n = 28$ and the number of cross correlation coefficients is equal to 378. The higher values of deviations $\delta_\alpha$ correspond to the higher structural divergence between compared sequences.

## RESULTS

**Nucleotide polymorphism and recombination:** Figure 1 shows a total of 92 polymorphic sites in a sample of 28 sequences of the ψ*Est-6* putative pseudogene: 62 sites in exon I (1 site involves a 3-bp deletion), 2 sites in the intron, 12 sites in exon II, and 16 sites in the 3′-flanking region. Two indel polymorphisms occur. A 3-bp deletion occurs in the F-775F strain and a 9-bp insertion occurs in the S-510S, S-501S, and S-5F strains (Figure 1). For the ψ*Est-6* coding region we detected 41 replacements (1 site involves a 9-bp insertion) and 33 synonymous polymorphic sites. We previously found 13 replacements and 23 synonymous polymorphic sites in the *Est-6* coding region (BALAKIREV *et al.* 2002). The ratio of replacement to synonymous polymorphic sites is 0.565 for the *Est-6* gene and more than twice that, 1.242, for the putative pseudogene. We detected 11 premature stop codons (all TGA) within the coding region of the putative pseudogene. The stop codons are generated by single mutations (positions 1068 and 1388) as well as by the insertion ACATTTGAT (Figure 1). The *mdg-3* retrotransposon insertion (5.2 kb) was detected within the intron of ψ*Est-6* in the S-438S *D. melanogaster* strain (data not shown). GAME and OAKESHOTT (1990) detected the same insertion previously in strain 12I-11.2 of *D. melanogaster*, which carried a null allele of ψ*Est-6*.

Table 1 shows estimates of nucleotide diversity for the putative pseudogene as well as for the *Est-6* gene and the flanking regions. The π value for the full sequence is 0.0084, which is within the range of values observed in other high-recombination gene regions in *D. melanogaster* (MORIYAMA and POWELL 1996). The π value is very similar in the 5′-flanking (0.0059) and *Est-6* regions

TABLE 1

**Nucleotide diversity and divergence of *Est-6*, *ψEst-6*, and flanking regions, in 28 strains of *D. melanogaster***

| | 5′-flanking region | *Est-6* | | | Intergenic region | *ψEst-6* | | | 3′-flanking region | Full sequence |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Syn | Nsyn | Total | | Syn | Nsyn | Total | | |
| $N$ | 1183 | | 1686 | | 193 | | 1700 | | 632 | 5394 |
| $S$ | 26 (8) | | 37 (12) | | 14 (7) | | 76 (14) | | 16 (3) | 171 (44) |
| $\pi$ | 0.0059 | 0.0152 | 0.0026 | 0.0057 | 0.0141 | 0.0268 | 0.0078 | 0.0122 | 0.0083 | 0.0084 |
| $\theta$ | 0.0060 | 0.0156 | 0.0027 | 0.0056 | 0.0186 | 0.0224 | 0.0080 | 0.0114 | 0.0065 | 0.0081 |
| $N_{\mathrm{hap}}$ | 13 | | 15 | | 6 | | 12 | | 6 | 23 |
| $k$ | 6.608 | | 9.542 | | 2.714 | | 20.534 | | 5.262 | 44.661 |
| $K$ | 0.0807 | 0.1469 | 0.0215 | 0.0495 | 0.0716 | 0.1393 | 0.0311 | 0.0546 | 0.0417 | 0.0572 |
| $Z_{\mathrm{nS}}$ | 0.2093 | | 0.2036 | | 0.2458 | | 0.4221 | | 0.4297 | 0.2477 |
| $\mathrm{LD}_{\mathrm{Fisher}}$ | 0.2646 | | 0.2703 | | 0.2418 | | 0.5596 | | 0.5750 | 0.3313 |

$N$ is the number of sites. $S$ is the number of polymorphic sites (the number of singleton polymorphic sites is in parentheses). $\pi$ is the average number of nucleotide differences per site among all pairs of sequences (Nei 1987, p. 256). $\theta$ is the average number of segregating nucleotide sites among all sequences based on the expected distribution of neutral variants in a panmictic population at equilibrium (Watterson 1975). $N_{\mathrm{hap}}$ is the number of haplotypes. $k$ is the average number of nucleotide differences. $K$ is the average proportion of nucleotide differences between *D. melanogaster* and *D. simulans*, corrected according to Jukes and Cantor (1969). $Z_{\mathrm{nS}}$ is a statistic of Kelly's (1997) test. $\mathrm{LD}_{\mathrm{Fisher}}$ is the proportion of significant linkage disequilibrium revealed by Fisher's exact test using all polymorphic sites. Syn, synonymous; Nsyn, nonsynonymous. The segregating sites associated with indels are excluded from the $\pi$, $\theta$, and $K$ calculations.

(0.0057), but significantly higher in the intergenic region (0.0141) and putative pseudogene (0.0122), and intermediate in the 3′-flanking region (0.0083). The level of synonymous variation is 0.0152 in the *Est-6* coding region but 0.0268 (1.76 times higher) in the putative pseudogene. The difference is more pronounced for nonsynonymous variation, which is 0.0026 in the *Est-6* gene and 0.0078 (3.0 times higher) in the putative pseudogene. This could indicate different degrees of selective constraint in the *Est-6* gene and the putative pseudogene. The level of silent polymorphism in the 3′-flanking region is 0.0083, but 0.0268 (3.2 times higher) in the putative pseudogene. These differences could again indicate differences in selective constraints. The level of silent divergence between *D. melanogaster* and *D. simulans* is similar for the *Est-6* gene (0.1469) and the putative pseudogene (0.1393), but lower in the 5′-flanking (0.0807) and 3′-flanking (0.0417) regions.

The method of Hudson and Kaplan (1985) reveals a minimum of 15 recombination events in the whole region (5394 bp) analyzed. The minimum number of recombination events is six for the *Est-6* gene but three for the putative pseudogene. There is a large difference (3300 times) in the value of the recombination estimator $C$ (Hudson 1987) obtained for the *Est-6* gene and the putative pseudogene (Table 2). The same tendency (but much less pronounced) is obtained using the permutation-based method of McVean *et al.* (2002) and also the method of Hey and Wakeley (1997) based on the number of pairs of sites with incongruent genealogical histories (Table 2).

Thus, there is two times more total nucleotide variability in the putative pseudogene (Table 1) but the recombination rate is at least two times higher in the *Est-6* gene (Table 2). The association in *ψEst-6* of a high level of nucleotide variation with low recombination is contrary to the well-documented positive relationship between within-species DNA variation and recombination rates (*e.g.*, Begun and Aquadro 1992).

**Haplotype structure:** Previously, Odgers *et al.* (1995)

TABLE 2

**Recombination estimates**

| | | $C$ | | | $\gamma$ | | | $\rho$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R_{\mathrm{m}}$ | Per gene | Per site | $C/\theta$ | Per gene | Per site | $\gamma/\theta$ | Per gene | Per site | $\rho/\theta$ |
| Full sequence | 15 | 1.0 | 0.0002 | 0.0247 | 40.828 | 0.0076 | 0.9383 | 9.619 | 0.0018 | 0.2222 |
| *Est-6* | 6 | 3.3 | 0.0020 | 0.3571 | 25.144 | 0.0149 | 2.6607 | 10.020 | 0.0059 | 1.0351 |
| *ψEst-6* | 3 | 0.001 | 0.0000 | — | 9.834 | 0.0058 | 0.5088 | 2.605 | 0.0015 | 0.1316 |

$C$, $\gamma$, and $\rho$ are estimates of the population recombination rate $4N_{\mathrm{e}}r$ ($N_{\mathrm{e}}$ is the effective population size and $r$ is the recombination rate per nucleotide site per generation) obtained by the methods of Hudson (1987), Hey and Wakeley (1997), and McVean *et al.* (2002), respectively. $R_{\mathrm{m}}$ is the minimum number of recombination events (Hudson and Kaplan 1985).

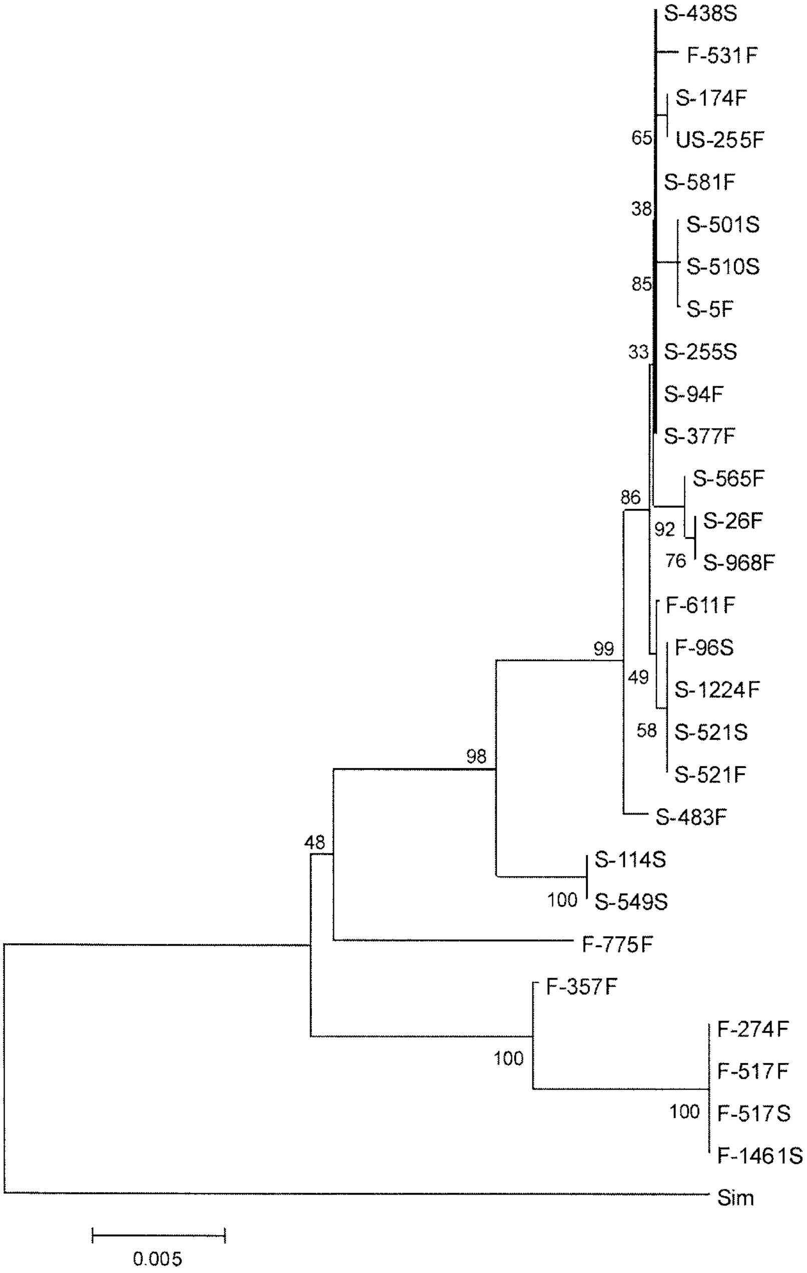

Figure 2.—Neighbor-joining tree of the ψ*Est-6* haplotypes of *D. melanogaster* based on Kimura's two-parameter distance. The numbers at the nodes are percentages of bootstrap probability values based on 10,000 replications.

and Balakirev *et al.* (1999, 2002) described two groups of haplotypes for both the 5′-flanking and *Est-6* coding regions in *D. melanogaster.* There are also two groups of haplotypes in the putative pseudogene region (Figures 1 and 2) that are labeled S or F according to the *Est-6* haplotype with which they are associated. Note, however, that lines F-531F, F-611F, and F-96S belong to the fast allozyme group of the *Est-6* gene (Balakirev *et al.* 2002) but they are in the slow allozyme group in the neighbor-joining tree of the putative pseudogene sequences (Figure 2). The average number of nucleotide differences ($K$) between the two haplotypes (excluding 3′-flanking region) is 20.534. The S group includes most haplotypes (22 out of 28), which are 3.69 times less variable ($\pi = 0.0026 \pm 0.0006$) than the six F haplotypes ($\pi = 0.0096 \pm 0.0042$). The permutation test of Hudson *et al.* (1992) is highly significant for the F and S haplotypes, $K^*_{st} = 0.3438$ ($P < 0.001$).

**Gene conversion:** The method of Sawyer (1989, 1999) detects gene conversion events within both the *Est-6* gene (5 regions in 10 sequences, $P = 0.0097$) and ψ*Est-6* (13 regions in all 28 sequences, $P = 0.0000$). The numbers of significant fragments are 14 for the *Est-6* gene (fragment length from 314 to 1183 bp, average 662 bp) and 85 for ψ*Est-6* (fragment length from 154 to 1052 bp, average 669 bp). Gene conversion events between *Est-6* and ψ*Est-6* are detected only in the protein alignment (involving a single region between amino acids 41 and 55, $P = 0.0102$). The number of significant fragment pairs showing intergenic conversion, which

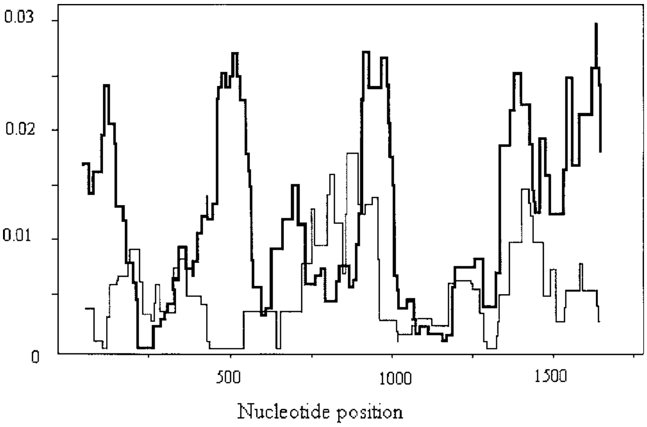

FIGURE 3.—Sliding-window plots of nucleotide diversity ($\pi$) in *Est-6* (thin line) and $\psi$*Est-6* (thick line). Window sizes are 100 nucleotides with 1-nucleotide increments.

involve 23 *Est-6* and 6 $\psi$*Est-6* sequences, is 138. Taken together, these results show that gene conversion has played an important role in the evolution of the β-*esterase* gene cluster.

**Sliding-window analysis:** Figure 3 shows the distribution of polymorphism along the *Est-6* (thin line) and $\psi$*Est-6* (thick line) sequences. There is a distinct peak in the *Est-6* sequences at 750–950, which includes the F/S replacement site (position 772). We detected this peak previously (Balakirev *et al.* 1999, 2002) in our 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). There are four (approximately equal) strong peaks in the $\psi$*Est-6* sequences at 50–200, 400–600, 850–1050, and 1300–1650. The putative pseudogene peaks are more acute than the *Est-6* gene peaks, have a regular distribution along the sequence (with an interval of 200–300 bp), and are not centered around the replacement polymorphisms (Figure 1).

We measure heterogeneity in the distribution of polymorphic sites along the $\psi$*Est-6* sequence and discordance between $\pi$ (within-*melanogaster* polymorphism) and $K$ (*melanogaster-simulans* divergence) by means of Goss and Lewontin's (1996) and McDonald's (1996, 1998) statistics and assess 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: Goss and Lewontin's (1996) interval length variance ($V_{IL}$) is 0.000164, $P > 0.05$, and modified interval length variance ($Q_{IL}$) is 0.000767, $P > 0.05$; McDonald's (1998) maximum sliding $G$ statistic is 6.9202, $P > 0.05$; the Kolmogorov-Smirnov statistic is 0.043733, $P > 0.05$. The tests are significant for the *Est-6* gene including the promoter region, but not for the *Est-6* coding region alone (Balakirev *et al.* 2002).

**Linkage disequilibrium:** We have calculated the $P$ value of Fisher's exact test in all pairwise comparisons of informative polymorphic sites. The numbers (and percentages) of pairwise comparisons that are significant are, for the whole region, 4235 out of 7626 (55.53%, 2.62% with the Bonferroni correction); for the *Est-6* gene, 151 out of 300 (50.33%, 18.33% with the correction); for the putative pseudogene, 1486 out of 1830 (81.20%, 14.81% with the correction); for the 3′-flanking region, 66 out of 78 (84.62%, 57.69% with the correction); and between *Est-6* and $\psi$*Est-6*, 927 of 1525 (60.79%, but none with the Bonferroni correction). The significant interlocus linkage disequilibria are caused by six divergent haplotypes, F-517S, F-517F, F-1461S, F-274F, F-357F, and F-775F, which have unique polymorphisms in both *Est-6* and $\psi$*Est-6*.

**Tests of neutrality:** In a previous study (Balakirev *et al.* 2002) we detected significant deviations from neutrality in the 5′-flanking and *Est-6* coding regions, using Kelly's $Z_{nS}$ (1997) and Wall's (1999) $B$ and $Q$ tests, based on linkage disequilibrium between segregating sites; both tests were significant with the population recombination rate ≥0.010 (Kelly's test) or without recombination (Wall's test). For $\psi$*Est-6*, Kelly's $Z_{nS}$ and Wall's $B$ and $Q$ values are even higher ($Z_{nS} = 0.422$; $B = 0.432$; $Q = 0.520$) than those for the *Est-6* gene and significant by coalescent simulations with the population recombination rate ≥0.005 ($Z_{nS}$ statistic) or without recombination ($B$ and $Q$ statistics).

**Entropy analysis:** We use this new type of analysis when seeking to ascertain the functionality of $\psi$*Est-6*. We have calculated the relevant characteristics for the exon-intron-exon sequences of *Est-6* and $\psi$*Est-6* before splicing and for exon-exon sequences after splicing. The examples of spectra for structure factor harmonics (see materials and methods, Equation 5) are illustrated in Figures 4 and 5, where the period $p$ is related to the ordinal number of structure factor harmonic $n$ as $p = M/n$. The high peaks at $p = 3$ ($n = 561$ for *Est-6* and $n = 563$ for $\psi$*Est-6*) are a distinctive feature of protein-coding regions (Fickett and Tung 1992) and are inherent to both genes and pseudogenes (Holste *et al.* 2000). The characteristics for the structural entropy averaged over the 28 lines of *D. melanogaster* are presented in Table 3. The values of integral spectral structural entropy per harmonic summed over all four types of nucleotides, $S/(M - 1)$, are equal to $-1.808$ ($r = 3.16$) for *Est-6*, $-1.821$ ($r = 3.45$) for spliced *Est-6*, $-1.718$ ($r = 0.73$) for $\psi$*Est-6*, and $-1.736$ ($r = 1.20$) for spliced $\psi$*Est-6*. The values obtained for $\psi$*Est-6* are not significantly different from random sequence, while the entropy of the *Est-6* gene is significantly higher than expected for random sequence (Table 3). These results demonstrate higher structural ordering in *Est-6* than in $\psi$*Est*-6, in accordance with expectations if $\psi$*Est*-6 is indeed a pseudogene. (Note the generally higher values of $r$, indicating higher structural ordering in the spliced genes, with the excep-

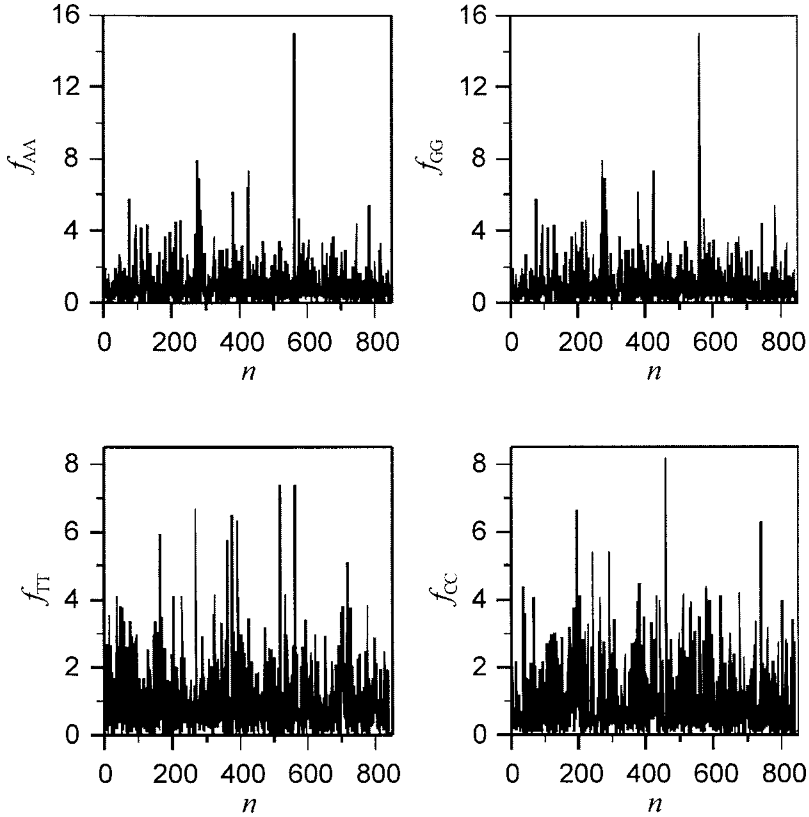

FIGURE 4.—The normalized structure factor spectra (see MATERIALS AND METHODS, Equations 1–5) for the unspliced *Est-6* gene. The high peaks at $n = 561$ correspond to three-periodicity ($p = 3$), which is a fundamental feature of protein-coding regions.

tion of a few cases due to structural coupling between exons and intron.)

The mean values of deviations $\delta_\alpha$ averaged over the set of 378 pairwise cross correlation coefficients corresponding to each gene are summarized in Table 4, while the examples of their distributions are presented in Figures 6 and 7. The values $\langle\delta\rangle$ in the last column in Table 4 are obtained by additional averaging of $\delta_\alpha$ over four types of nucleotides. The insertions/deletions are present only in ψ*Est-6* and produce the main contribution to deviations $\delta_\alpha$. Comparison of structural divergence in ψ*Est-6* was also performed upon equalizing the lengths of the sequences by removing insertions/deletions. As seen in Table 4, even after equalizing, the structural divergence remains distinctly higher in ψ*Est-6* than in *Est-6*, as expected if ψ*Est-6* is not a functional gene. Besides that, it is worth noting the correlation between the deviations $\delta_\alpha$ and the heights of peaks for $f_{\alpha\alpha}$ at $p = 3$ (see Equation 5 as well as Figures 4 and 5): the higher the peaks at $p = 3$ the smaller $\delta_\alpha$ and the narrower their distributions (see Figures 6 and 7 as well as Table 4).

## DISCUSSION

**Pseudogenes in Drosophila and other organisms:** Relative to what is known in other organisms, especially vertebrates (MIGHELL *et al.* 2000), pseudogenes are not common in Drosophila (POWELL 1997). Moreover, sequence evolution in many of the pseudogenes detected in Drosophila has indications of some functional constraint, including lower than expected levels of intraspecific variability and interspecific divergence, significant heterogeneity of nucleotide variability and divergence along the sequences, higher rate of substitution at synonymous nucleotide positions, conservation of important functional regions, transcriptional activity, and codon bias (JEFFS and ASHBURNER 1991; CURRIE and SULLIVAN 1994; JEFFS *et al.* 1994; SULLIVAN *et al.* 1994; PRITCHARD and SCHAEFFER 1997; RAMOS-ONSINS and AGUADÉ 1998). Two *Adh* Drosophila genes originally identified as pseudogenes (FISHER and MANIATIS 1985; JEFFS and ASHBURNER 1991) were later considered to be novel functional genes (LONG and LANGLEY 1993; BEGUN 1997).

The unusual patterns of pseudogene evolution suggesting some functional constraints have also been revealed in other organisms. Moreover, it was shown that a small number of detrimental alterations is a common feature of pseudogenes; there are many examples of extremely conserved pseudogene sequences exhibiting 90% and higher homology with functional counterparts (for a review, see VANIN 1985; WILDE 1986; MIGHELL *et al.* 2000). The extent of similarity between a pseudogene and its functional counterpart could be sharply nonuniform along the sequence (*e.g.*, SUDO *et al.* 1990; MATTERS and GOODENOUGH 1992; JOHN *et al.* 1996).

By definition, pseudogenes should be transcriptionally and translationally silent; however, nonfunctional

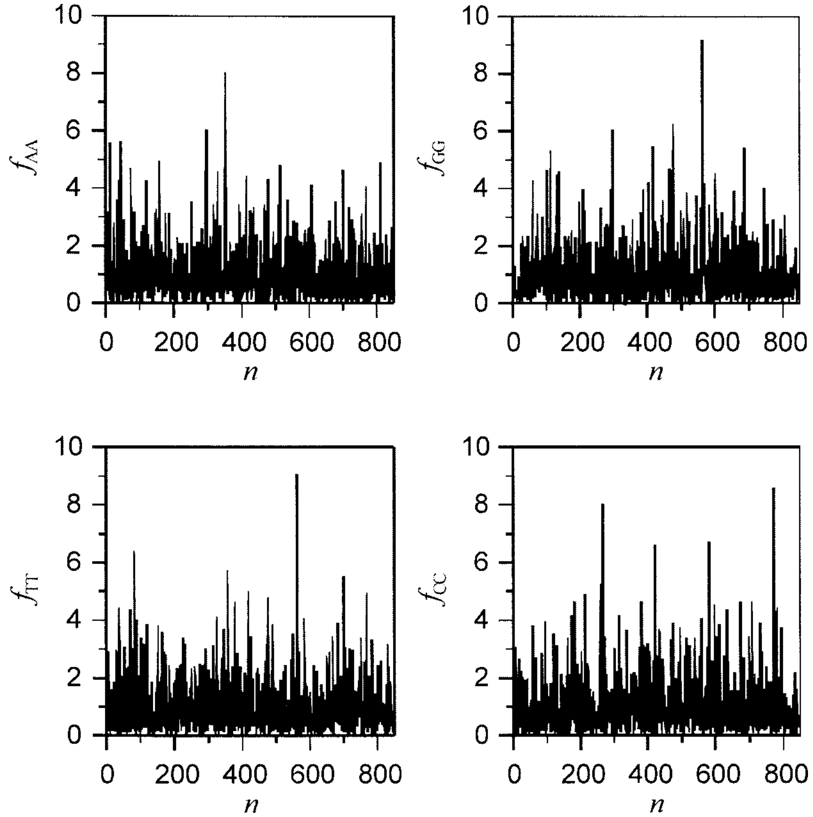

FIGURE 5.—The normalized structure factor spectra (see MATERIALS AND METHODS, Equations 1–5) for the unspliced *ψEst-6* putative pseudogene. The high peaks at $n = 563$ correspond to three-periodicity ($p = 3$), which is a fundamental feature of protein-coding regions.

(or functional only in some cases) transcripts of many pseudogenes have been described (*e.g.*, FOTAKI and IATROU 1988; SORGE *et al.* 1990; NGUYEN *et al.* 1991; ZHOU *et al.* 1992; CRISTIANO *et al.* 1993; BARD *et al.* 1995; FÜRBAB and VANSELOW 1995; for a review and further reference, see VANIN 1985; WILDE 1986; MIGHELL *et al.* 2000). In some cases, the translation activity was shown *in vivo* (MCCARREY and THOMAS 1987; BRISTOW *et al.* 1993) and *in vitro* (MISRA-PRESS *et al.* 1994). SORGE *et al.* (1990) also detected considerable subject-to-subject variation in the relative amounts of transcripts derived from gene and pseudogene that could be due to polymorphisms in the promoter regions of the pseudogene or to polymorphisms affecting stability of pseudogene-derived mRNA. This situation is reminiscent of the *ψEst-6* putative pseudogene, for which stop codons were detected only in some lines of *D. melanogaster*.

**Pattern of recombination:** Disruption of homologous sequences by insertion or nucleotide polymorphisms can significantly reduce recombination frequencies. Even individual nucleotide substitutions have been shown to affect recombination (*e.g.*, SELVA *et al.* 1995; LUKACSOVICH and WALDMAN 1999). In comparison with recombination between alleles lacking large insertion polymorphisms, recombination within the maize *a1* gene (XU *et al.* 1995) is inhibited when one allele with a transposon insertion is paired with a second allele lacking an insertion. In this respect, we note that the insertion of the *mdg-3* retrotransposon within the intron of the *ψEst-6* putative pseudogene has been detected in one *D. melanogaster* strain studied here (S-438S) and also in strain 12I-11.2 analyzed by GAME and OAKESHOTT (1990).

Recombination is not a random process; recombination hotspots caused by specific initiating sequences are reported for many gene systems. The unprecedented evolutionary stability of simple repeats promoting recombination in the expressed mammalian *MHC-DRB* genes was detected in some specific genome locations (review in SCHWAIGER and EPPLEN 1995). But in mammalian *DRB* pseudogenes, the simple repeat stretch seems gradually to reduce its characteristic pattern in the evolutionary course (LARHAMMAR *et al.* 1985; RIESS *et al.* 1990; SCHWAIGER and EPPLEN 1995). BLISKOVSKII *et al.* (1993) described the structure of the human *son* processed pseudogene, which has a 96% homology with the *son* functional gene. Despite the high sequence homology, the *son* pseudogene lacks five monomers of the perfect tandem repeat area, which, it has been suggested, are associated with the initiation of recombination processes (JEFFREYS *et al.* 1985).

In several yeast and maize genes, the sequence signals initiating recombination often occur within the promoter but not within the gene itself (WHITE *et al.* 1993; FAN *et al.* 1995; XU *et al.* 1995). The promoter of the *ψEst-6* putative pseudogene is limited to 193 bp consisting of the intergenic region between *Est-6* and *ψEst-6*. It might be that the obvious reduction in recombination could be connected with the promoter truncation of *ψEst-6*.

TABLE 3

**The characteristics for the structural entropy averaged over 28 lines of *Drosophila melanogaster***

| | Nucleotides | | | |
|---|---|---|---|---|
| Parameters | A | G | T | C |
| | | *Est-6* | | |
| $S_\alpha/(M-1)$ | −0.458 | −0.468 | −0.437 | −0.445 |
| $S_{\alpha,\mathrm{rel}}$ | 0.083 | 0.108 | 0.034 | 0.053 |
| $r_\alpha$ | 1.88 | 2.46 | 0.78 | 1.20 |
| | | Spliced *Est-6* | | |
| $S_\alpha/(M-1)$ | −0.446 | −0.475 | −0.452 | −0.448 |
| $S_{\alpha,\mathrm{rel}}$ | 0.054 | 0.124 | 0.069 | 0.060 |
| $r_\alpha$ | 1.22 | 2.77 | 1.55 | 1.35 |
| | | *ψEst-6* | | |
| $S_\alpha/(M-1)$ | −0.433 | −0.429 | −0.407 | −0.448 |
| $S_{\alpha,\mathrm{rel}}$ | 0.024 | 0.015 | −0.036 | 0.061 |
| $r_\alpha$ | 0.56 | 0.34 | −0.83 | 1.38 |
| | | Spliced *ψEst-6* | | |
| $S_\alpha/(M-1)$ | −0.446 | −0.452 | −0.415 | −0.423 |
| $S_{\alpha,\mathrm{rel}}$ | 0.055 | 0.069 | −0.017 | −0.001 |
| $r_\alpha$ | 1.24 | 1.55 | −0.39 | −0.008 |

The sequences promoting recombination could also be eroded within *ψEst-6* due to stochastic accumulation of mutations as in the case of *HLA-DRB* (Larhammar *et al.* 1985; Riess *et al.* 1990; Schwaiger and Epplen 1995) and *son* (Bliskovskii *et al.* 1993) pseudogenes.

**Pseudogene function:** A possible role for pseudogenes in development as a source of the intracellular inhibitors was suggested by McCarrey and Riggs (1986). It has been also suggested that pseudogenes may have regulatory roles for the genes from which they have been derived (Fotaki and Iatrou 1988; Inouye 1988; Zhou *et al.* 1992; Troyanovsky and Leube 1994). Healy *et al.* (1996) have shown that 3′ sequences within the *ψEst-6* transcription unit contain elements that modulate the expression of *Est-6*.

A functional role has been proposed and, in some cases clearly brought out, for pseudogenes in the diversity of vertebrate immune response (*e.g.*, Reynaud *et al.* 1987; Knight 1992; Vargas-Madrazo *et al.* 1995; Sayegh *et al.* 1999). The immunoglobulin gene diversity is generated by somatic gene conversion events in which sequences derived from pseudogenes are integrated into functional germ-line genes. Gene conversion events have been also reported in several bacterial pathogens as a mechanism for generating antigenic variation (of sequence diversity in the expressed genes; *e.g.*, Thon *et al.* 1989; Zhang *et al.* 1997; Noormohammadi *et al.* 2000; Brayton *et al.* 2001). It has been suggested that human olfactory receptor (OR) pseudogenes might be important for the generation and maintenance of diversity (Glusman *et al.* 2000). While OR pseudogenes have lost coding function, they are apparently under new evolutionary constraints; OR pseudogenes adopt noncoding functions as CpG islands (Glusman *et al.* 2000), enhancers (Buettner *et al.* 1998), and matrix attachment regions (Gimelbrant and McClintock 1997). While pseudogenes are generally defined as nonfunctional, the above examples serve to challenge this characterization.

**Conclusions:** We have detected some contrasting characteristics of nucleotide variation in the *Est-6* gene and the *ψEst-6* putative pseudogene. The level of the total nucleotide variation is 2.1 times higher in *ψEst-6* than in *Est-6*. The population recombination rate is at least 2.6 times lower in *ψEst-6* than in *Est-6*. As a consequence, linkage disequilibrium is more pronounced in *ψEst-6* than in *Est-6*. The haplotype structure of *ψEst-6* is dimorphic. However, the divergent sequences of *ψEst-6* are not perfectly associated with *Est-6* allozyme variation. Some of the detected features of *ψEst-6* indicate that it could be a pseudogene: 11 premature stop codons out of 28 strains are hardly compatible with functionality of the encoded protein. The level of nonsynonymous variation is 3.0 times higher in *ψEst-6* than in *Est-6*. The results of the structural entropy analysis reveal a lower structural regularity and a higher structural divergence for *ψEst-6*, in accordance with the expectations provided it is a pseudogene or nonfunctional gene. On the other hand, it has been shown that the gene is expressed (Collet *et al.* 1990) and some alleles of *ψEst-6* produce

TABLE 4

**The mean divergences of cross correlations $\delta_\alpha$ within 28 lines of *Drosophila melanogaster***

| | Nucleotides | | | | |
|---|---|---|---|---|---|
| Genes | A | G | T | C | $\langle\delta\rangle$ |
| *Est-6* | 0.008 | 0.008 | 0.020 | 0.018 | 0.014 |
| Spliced *Est-6* | 0.008 | 0.009 | 0.019 | 0.018 | 0.013 |
| *ψEst-6* | 0.114 | 0.097 | 0.106 | 0.101 | 0.105 |
| Spliced *ψEst-6* | 0.101 | 0.085 | 0.095 | 0.097 | 0.095 |
| Equalized *ψEst-6* | 0.033 | 0.026 | 0.027 | 0.034 | 0.030 |
| Equalized spliced *ψEst-6* | 0.032 | 0.024 | 0.028 | 0.032 | 0.029 |

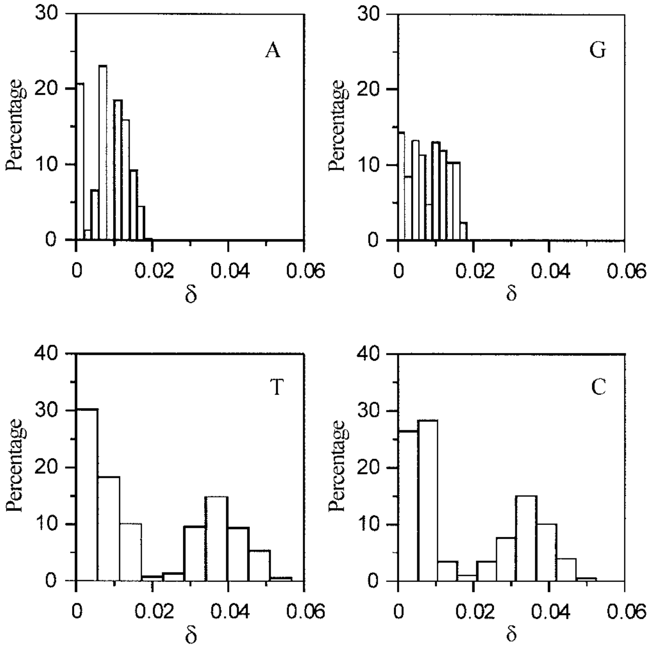

FIGURE 6.—Histograms for deviations $\delta_\alpha$ (see MATERIALS AND METHODS, Equations 11–13) for *Est-6* in 28 lines of *D. melanogaster*. Higher values of $\delta_\alpha$ correspond to higher structural divergence between compared sequences.

a catalytically active esterase (DUMANCIC *et al.* 1997). On the basis of the results we have presented and the review of the literature about pseudogene nucleotide variation, we conclude that there is not a sharp division between genes and pseudogenes with respect to function. A pseudogene may lose some specific function but retain or acquire another, which may not be simply recognizable. There are many examples of functional or "active" pseudogenes, a statement that would amount to a genetic oxymoron, if pseudogenes are defined as nonfunctional. Taking into account that the function of ψ*Est-6* is not known (but could exist and may be discovered in the future), we suggest that the term potogene be used for ψ*Est-6*, following the terminology of BROSIUS and GOULD (1992), reflecting that the status of this gene is not certain at present. These authors pointed out that the products of gene duplication, including those that become pseudogenes, may eventually acquire distinctive functions, and thus might be called *poto*genes to call attention to their *pot*entiality for becoming new genes or acquiring new functions. The distinctive features of ψ*Est-6*, including its pattern of population variation, suggest that it may already have some functional role (for instance, as a reservoir of sequences, which can recombine with the expressed *Est-6* gene), but its designation as a potogene would imply that the function is not known and far from confirmed, although the potentiality exists.

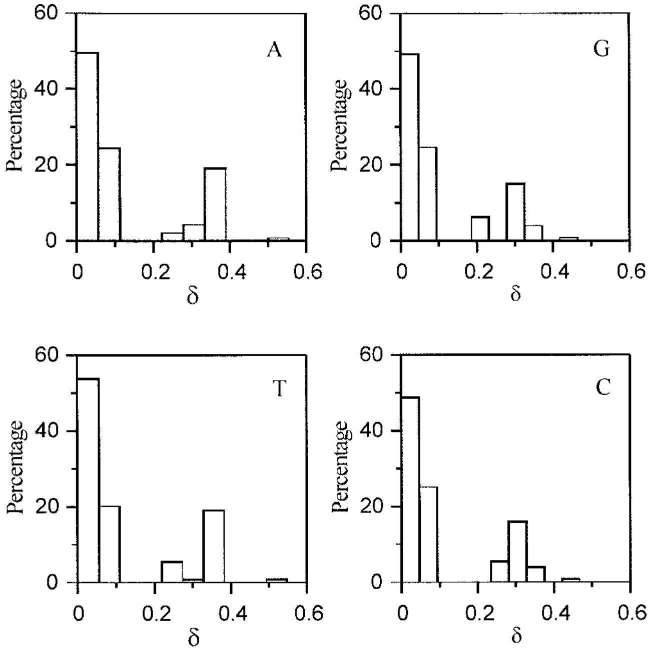

FIGURE 7.—Histograms for deviations $\delta_\alpha$ (see MATERIALS AND METHODS, Equations 11–13) for ψ*Est-6* in 28 lines of *D. melanogaster*. Higher values of $\delta_\alpha$ correspond to higher structural divergence between compared sequences.

We are grateful to S. A. Sawyer, 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 De-Filippis, Martina Zurovkova, and Carlos Márquez for encouragement and help; and W. M. Fitch, B. Gaut, R. R. Hudson, A. Long, and two anonymous reviewers for detailed and valuable comments. This work is supported by National Insititutes of Health grant GM42397 to F. J. Ayala.

## LITERATURE CITED

BALAKIREV, E. S., and F. J. AYALA, 1996 Is esterase-P encoded by a cryptic pseudogene in *Drosophila melanogaster*? Genetics **144:** 1511–1518.

BALAKIREV, E. S., E. I. BALAKIREV, F. RODRIGUEZ-TRELLES and F. J. AYALA, 1999 Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. Genetics **153:** 1357–1369.

BALAKIREV, E. S., E. I. BALAKIREV and F. J. AYALA, 2002 Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: contrasting patterns of DNA variability in adjacent functional regions. Gene **288:** 167–177.

BARD, J. A., S. P. NAWOSCHIK, B. F. O'DOWD, S. R. GEORGE, T. A. BRANCHEK *et al.*, 1995 The human serotonin 5-hydroxytryptamine$_{1D}$ receptor pseudogene is transcribed. Gene **153:** 295–296.

BEGUN, D., 1997 Origin and evolution of a new gene descended from *alcohol dehydrogenase* in Drosophila. Genetics **145:** 375–382.

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

BLISKOVSKII, V. V., A. V. KIRILLOV, K. S. SPIRIN, V. M. ZAKHAREV and I. M. CHUMAKOV, 1993 *Son* pseudogene does not contain five repeated elements of the area of perfect tandem repeats present in the homologous *son* gene sequence. Mol. Biol. **27:** 61–68.

BRAYTON, K. A., D. P. KNOWLES, T. C. MCGUIRE and G. H. PALMER, 2001 Efficient use of a small genome to generate antigenic diversity in tick-borne ehrlichial pathogens. Proc. Natl. Acad. Sci. USA **98:** 4130–4135.

BRISTOW, J., S. E. GITELMAN, M. K. TEE, B. STAELS and W. L. MILLER, 1993 Abundant adrenal-specific transcription of the human P450c21A "pseudogene". J. Biol. Chem. **268:** 12919–12924.

BROSIUS, J., and S. J. GOULD, 1992 On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk" DNA. Proc. Natl. Acad. Sci. USA **89:** 10706–10710.

Buettner, J. A., G. Glusman, N. Ben-Arie, P. Ramos, D. Lancet *et al.*, 1998 Organization and evolution of olfactory receptor genes on human chromosome 11. Genomics **53:** 56–68.

Chechetkin, V. R., and V. V. Lobzin, 1996 Levels of ordering in coding and non-coding regions of DNA sequences. Phys. Lett. A **222:** 354–360.

Chechetkin, V. R., and V. V. Lobzin, 1998 Study of correlations in segmented DNA sequences: application to structural coupling between exons and introns. J. Theor. Biol. **190:** 69–83.

Chechetkin, V. R., and A. Y. Turygin, 1994 On the spectral criteria of disorder in non-periodic sequences: application to inflation models, symbolic dynamics and DNA sequences. J. Phys. A Math. Gen. **27:** 4875–4898.

Chechetkin, V. R., and A. Y. Turygin, 1996 Study of correlations in DNA sequences. J. Theor. Biol. **178:** 205–217.

Chechetkin, V. R., L. A. Knizhnikova and A. Y. Turygin, 1994 Three-quasiperiodicity, mutual correlations, ordering and long-range modulations in genomic nucleotide sequences for viruses. J. Biomol. Struct. Dyn. **12:** 271–299.

Collet, C., K. M. Nielsen, R. J. Russell, M. Karl, J. G. Oakeshott *et al.*, 1990 Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. Mol. Biol. Evol. **7:** 9–28.

Cooke, P. H., and J. G. Oakeshott, 1989 Amino acid polymorphisms for esterase-6 in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 1426–1430.

Cristiano, R. J., S. J. Giordano and A. W. Steggles, 1993 The isolation and characterization of the bovine cytochrome $b_5$ gene, and a transcribed pseudogene. Genomics **17:** 348–354.

Currie, P. D., and D. T. Sullivan, 1994 Structure, expression and duplication of genes which encode phosphoglyceromutase of *Drosophila melanogaster*. Genetics **138:** 352–363.

Dumancic, M. M., J. G. Oakeshott, R. J. Russell and M. J. Healy, 1997 Characterization of the *EstP* protein in *Drosophila melanogaster* and its conservation in Drosophilids. Biochem. Genet. **35:** 251–271.

Fan, Q., F. Xu and T. Petes, 1995 Meiosis-specific double-strand breaks at the *HIS4* recombination hot spot in the yeast *Saccharomyces cerevisiae* control in *cis* and *trans*. Mol. Cell. Biol. **15:** 1679–1688.

Fickett, J. W., and C.-S. Tung, 1992 Assessment of protein coding measures. Nucleic Acids Res. **20:** 6441–6450.

Filatov, D. A., and D. Charlesworth, 1999 DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. Genetics **153:** 1423–1434.

Fisher, J. A., and T. Maniatis, 1985 Structure and transcription of the *Drosophila mulleri* alcohol dehydrogenase genes. Nucleic Acids Res. **13:** 6899–6917.

Fotaki, M. E., and K. Iatrou, 1988 Identification of a transcriptionally active pseudogene in the chorion locus of the silkmoth *Bombyx mori*. Regional sequence conservation and biological function. J. Mol. Biol. **203:** 849–860.

Fürbab, R., and J. Vanselow, 1995 An aromatase pseudogene is transcribed in the bovine placenta. Gene **154:** 287–291.

Game, A. Y., and J. G. Oakeshott, 1990 Associations between restriction site polymorphism and enzyme activity variation for esterase 6 in *Drosophila melanogaster*. Genetics **126:** 1021–1031.

Gimelbrant, A. A., and T. S. McClintock, 1997 A nuclear matrix attachment region is highly homologous to a conserved domain of olfactory receptors. J. Mol. Neurosci. **9:** 61–63.

Glusman, G., A. Sosinsky, E. Ben-Asher, N. Avidan, D. Sonkin *et al.*, 2000 Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. Genomics **63:** 227–245.

Goss, P. J. E., and R. C. Lewontin, 1996 Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589–602.

Gromko, M. H., D. F. Gilbert and R. C. Richmond, 1984 Sperm transfer and use in the multiple mating system of Drosophila, pp. 371–426 in *Sperm Competition and the Evolution of Animal Mating Systems*, edited by R. L. Smith. Academic Press, New York.

Hasson, E., and W. F. Eanes, 1996 Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics **144:** 1565–1575.

Healy, M. J., M. M. Dumancic and J. G. Oakeshott, 1991 Biochemical and physiological studies of soluble esterases from *Drosophila melanogaster*. Biochem. Genet. **29:** 365–388.

Healy, M. J., M. M. Dumancic, A. Cao and J. G. Oakeshott, 1996 Localization of sequences regulating ancestral and acquired sites of esterase 6 activity in *Drosophila melanogaster*. Mol. Biol. Evol. **13:** 784–797.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Holste, D., O. Weiss, I. Grosse and H. Herzel, 2000 Are noncoding sequences of *Rickettsia prowazekii* remnants of "neutralized" genes? J. Mol. Evol. **51:** 353–362.

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Biol. **7:** 1–44.

Hudson, R. R., and N. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

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

Hudson, R. R., D. Boos and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

Inouye, M., 1988 Antisense RNA: its functions and applications in gene regulation—a review. Gene **72:** 25–34.

Jeffreys, A. J., V. Wilson and S. L. Thein, 1985 Hypervariable minisatellite regions in human DNA. Nature **314:** 67–73.

Jeffs, P. S., and M. Ashburner, 1991 Processed pseudogenes in Drosophila. Proc. R. Soc. Lond. Ser. B **244:** 151–159.

Jeffs, P. S., E. C. Holmes and M. Ashburner, 1994 The molecular evolution of the Alcohol dehydrogenase and Alcohol dehydrogenase-related genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **11:** 287–304.

John, T. R., J. J. Smith and I. I. Kaiser, 1996 A phospholipase $A_2$-like pseudogene retaining the highly conserved introns of Mojave toxin and other snake venom group II $PLA_2$s, but having different exons. DNA Cell Biol. **15:** 661–668.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–120 in *Mammalian Protein Metabolism*, edited by H. M. Munro. Academic Press, New York.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Knight, K. L., 1992 Restricted $V_H$ usage and generation of antibody diversity in rabbit. Annu. Rev. Immunol. **10:** 593–616.

Larhammar, D., B. Servenius, L. Rask and P. A. Peterson, 1985 Characterization of an HLA DR-beta pseudogene. Proc. Natl. Acad. Sci. USA **82:** 1475–1479.

Lobzin, V. V., and V. R. Chechetkin, 2000 Order and correlations in genomic DNA sequences: the spectral approach. Phys. Usp. **43:** 55–78.

Long, M. Y., and C. H. Langley, 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science **260:** 91–95.

Lukacsovich, T., and A. S. Waldman, 1999 Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. Genetics **151:** 1559–1568.

Marple, Jr., S. L., 1987 *Digital Spectral Analysis With Applications*. Prentice Hall, Englewood Cliffs, NJ.

Matters, G. L., and U. W. Goodenough, 1992 A gene/pseudogene tandem duplication encodes a cysteine-rich protein expressed during zygote development in *Chlamydomonas reinhardtii*. Mol. Gen. Genet. **232:** 81–88.

McCarrey, J. R., and A. D. Riggs, 1986 Determinator-inhibitor pairs as a mechanism for threshold setting in development: a possible function for pseudogenes. Proc. Natl. Acad. Sci. USA **83:** 679–683.

McCarrey, J. R., and K. Thomas, 1987 Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature **326:** 501–505.

McDonald, J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **13:** 253–260.

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.

McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

Mighell, A. J., N. R. Smith, P. A. Robinson and A. F. Markham, 2000 Vertebrate pseudogenes. FEBS Lett. **468:** 109–114.

Misra-Press, A., N. E. Cooke and S. A. Liebhaber, 1994 Complex alternative splicing partially inactivates the human chorionic somatomammotropin-like (hCS-L) gene. J. Biol. Chem. **269:** 23220–23229.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

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

Nguyen, T., R. Sunahara, A. Marchese, H. H. M. Van Tol, P. Seeman *et al.*, 1991 Transcription of a human dopamine pseudogene. Biochem. Biophys. Res. Commun. **181:** 16–21.

Noormohammadi, A. H., P. F. Markham, A. Kanci, K. G. Whithear and G. F. Browning, 2000 A novel mechanism for control of antigenic variation in the haemagglutinin gene family of *Mycoplasma synoviae.* Mol. Microbiol. **35:** 911–923.

Oakeshott, J. G., C. Collet, R. Phillis, K. M. Nielsen, R. J. Russell *et al.*, 1987 Molecular cloning and characterization of esterase 6, a serine hydrolase from Drosophila. Proc. Natl. Acad. Sci. USA **84:** 3359–3363.

Oakeshott, J. G., E. A. van Papenrecht, T. M. Boyce, M. J. Healy and R. J. Russell, 1993 Evolutionary genetics of Drosophila esterases. Genetica **90:** 239–268.

Odgers, W. A., M. J. Healy and J. G. Oakeshott, 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.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, Oxford/New York.

Pritchard, J. K., and S. W. Schaeffer, 1997 Polymorphism and divergence at a *Drosophila* pseudogene locus. Genetics **147:** 199–208.

Procunier, W. S., J. J. Smith and R. C. Richmond, 1991 Physical mapping of the *esterase-6* locus of *Drosophila melanogaster.* Genetica **84:** 203–208.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Reynaud, C. A., V. Anquez, A. Dahan and J. C. Weill, 1987 A hyperconversion mechanism generates the chicken preimmune light chain repertoire. Cell **48:** 379–388.

Richmond, R. C., D. G. Gilbert, K. B. Sheehan, M. H. Gromko and F. M. Butterworth, 1980 Esterase 6 and reproduction in *Drosophila melanogaster.* Science **207:** 1483–1485.

Richmond, R. C., K. M. Nielsen, J. P. Brady and E. M. Snella, 1990 Physiology, biochemistry and molecular biology of the *Est-6* locus in *Drosophila melanogaster*, pp. 273–292 in *Ecological and Evolutionary Genetics of Drosophila*, edited by J. S. F. Barker, W. T. Starmer and R. J. MacIntyre. Plenum Press, New York.

Riess, O., C. Kammerbauer, L. Roewer, V. Steimle, A. Andreas *et al.*, 1990 Hypervariability of intronic simple $(gt)_n(ga)_m$ repeats in *HLA-DRB* genes. Immunogenetics **32:** 110–116.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

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

Sawyer, S. A., 1999 *GENECONV*: a computer package for the statistical detection of gene conversion. Department of Mathematics, Washington University, St. Louis.

Sayegh, C. E., G. Drury and J. H. Ratcliffe, 1999 Efficient antibody diversification by gene conversion *in vivo* in the absence of selection for V(D)J-encoded determinants. EMBO J. **18:** 6319–6328.

Schwaiger, F.-W., and J. T. Epplen, 1995 Exonic *MHC-DRB* polymorphisms and intronic simple repeat sequence: *Janus*' faces of DNA sequence evolution. Immunol. Rev. **143:** 199–224.

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

Selva, E. M., L. New, G. F. Crouse and R. S. Lahue, 1995 Mismatch correction acts as a barrier to homologous recombination in *Saccharomyces cerevisiae.* Genetics **139:** 1175–1188.

Sorge, J., E. Gross, C. West and E. Beutler, 1990 High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. J. Clin. Invest. **86:** 1137–1141.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

Sudo, K., M. Maekawa, M. M. Luedemann, L. L. Deaven and S. S.-L. Li, 1990 Human lactate dehydrogenase-B processed pseudogene: nucleotide sequence analysis and assignment to the X-chromosome. Biochem. Biophys. Res. Com. **171:** 67–74.

Sullivan, D. T., W. T. Starmer, S. W. Curtiss, M. Menotti-Raymond and J. Yum, 1994 Unusual molecular evolution of an *Adh* pseudogene in Drosophila. Mol. Biol. Evol. **11:** 443–458.

Thon, G., T. Baltz and H. Eisen, 1989 Antigenic diversity by the recombination of pseudogenes. Genes Dev. **3:** 1247–1254.

Troyanovsky, S. M., and R. E. Leube, 1994 Activation of the silent human cytokeratin 17 pseudogene-promoter region by cryptic enhancer elements of the cytokeratin 17 gene. Eur. J. Biochem. **223:** 61–69.

Vanin, E., 1985 Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. **19:** 253–272.

Vargas-Madrazo, E., J. C. Almagro and F. Lara-Ochoa, 1995 Structural repertoire in $V_H$ pseudogenes of immunoglobulins: comparison with human germline genes and human amino acid sequences. J. Mol. Biol. **246:** 74–81.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

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

White, M. A., M. Dominska and T. D. Petes, 1993 Transcription factors are required for the meiotic recombination hotspot at the *HIS4* locus in *Saccharomyces cerevisiae.* Proc. Natl. Acad. Sci. USA **90:** 6621–6625.

Wilde, C. D., 1986 Pseudogenes. CRC Crit. Rev. Biochem. **19:** 323–352.

Xu, X., A.-P. Hsia, L. Zhang, B. J. Nikolau and P. S. Schnable, 1995 Meiotic recombination break points resolve at high rates at the 5′ end of a maize coding sequence. Plant Cell **7:** 2151–2161.

Zhang, J.-R., J. M. Hardham, A. G. Barbour and S. J. Norris, 1997 Antigenic variation in lyme disease Borrelia by promiscuous recombination of VMP-like sequence cassettes. Cell **89:** 275–285.

Zhou, B.-S., D. R. Beidler and Y.-C. Cheng, 1992 Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. Cancer Res. **52:** 4280–4285.

Communicating editor: N. Takahata