EXCESS POLYMORPHISM AT THE *ADH* LOCUS IN *DROSOPHILA MELANOGASTER*

MARTIN E. KREITMAN[1] AND MONTSERRAT AGUADÉ[2]

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received January 1, 1986
Revised copy accepted May 5, 1986

ABSTRACT

The evolutionary history of a region of DNA encompassing the *Adh* locus is studied by comparing patterns of variation in *Drosophila melanogaster* and its sibling species, *D. simulans*. An unexpectedly high level of silent polymorphism in the *Adh* coding region relative to the 5′ and 3′ flanking regions in *D. melanogaster* is revealed by a populational survey of restriction polymorphism using a four-cutter filter hybridization technique as well as by direct sequence comparisons. In both of these studies, a region of the *Adh* gene encompassing the three coding exons exhibits a frequency of polymorphism equal to that of a 4-kb 5′ flanking region. In contrast, an interspecific sequence comparison shows a two-fold higher level of divergence in the 5′ flanking sequence compared to the structural locus. Analysis of the patterns of variation suggest an excess of polymorphism within the *D. melanogaster Adh* locus, rather than lack of polymorphism in the 5′ flanking region. An approach is outlined for testing neutral theory predictions about patterns of variation within and between species. This approach indicates that the observed patterns of variation are incompatible with an infinite site neutral model.

MANY evolutionary models contain predictions about the relationship between standing levels of genetic variation in natural populations and substitution rates as measured by the extent of divergence between species. Essentially all neutral theory models, for example, predict a direct proportionality between the two variables, and certain selective models also predict a positive relationship. However, there is also a class of well-considered selective models for which a strong positive correlation is not expected. This includes overdominance models in which polymorphism is maintained by heterozygote advantage, but in which the substitution rate is governed by some other mechanism (*e.g.*, positive selection or neutral drift).

Given that evolutionary models can be classified according to their prediction about the relationship between standing variation and substitution rate, it is worthwhile to consider whether nucleotide data can be used to evaluate this relationship. It should be immediately clear that such a test cannot be based

[1] Current address: Department of Biology, Princeton University, Princeton, New Jersey 08544.
[2] Current address: Departament de Genetica, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain.

on a single estimate of polymorphism and a single estimate of divergence: a null hypothesis must also be generated from the data.

This requirement can be easily met by comparing levels of variation and divergence across two or more regions of DNA. A test can then be constructed to determine whether estimates of the levels of variation in different regions of DNA are statistically independent of estimates of the substitution rates for the same regions. This comparison asks whether a pattern of variation based on allelic differences within a species is consistent with a pattern of fixed differences between species.

In order to determine the significance level for any observed comparison, some assumption must first be made about how mutations are expected to be distributed along the DNA. One simple model assumes that mutations occur randomly along a sequence and that each site evolves independently (*e.g.*, free recombination between sites). Indeed, this is the assumption underlying KIMURA'S infinite site neutral model (KIMURA 1969). Under this assumption, mutations are expected to be Poisson distributed, and the significance levels for tests of independence are easily obtained from parametric distributions (*e.g.*, χ^2). This implies that the infinite site neutral model as well as many others can, indeed, be evaluated from appropriate nucleotide data.

Utilizing this approach, we investigate here the evolutionary properties of noncoding and coding regions of DNA encompassing the *Adh* locus in *D. melanogaster* and its sibling species, *D. simulans*. First, we analyze the DNA sequences of a 4.5-kb region upstream from the *Adh* locus for two *D. melanogaster* alleles and compare this with sequence differences between the same alleles for the *Adh* structural locus. Then, we generalize these results to natural populations of *D. melanogaster* by comparing heterozygosity estimates from restriction polymorphisms for the same regions of DNA, using a technique that reveals four-cutter restriction polymorphisms. Finally, we compare these results with that of a sequence comparison for the *Adh* locus and its 5′ region between *D. melanogaster* and *D. simulans*. We show from intraspecific comparisons that there is a significantly higher level of silent variation within the structural locus than in the 5′ or 3′ flanking regions. Interspecific sequence comparisons show that this difference cannot be explained by differential constraint in the two regions. This leads us to reject a neutral model to account for the apparent excess silent polymorphism in the *Adh* locus.

MATERIALS AND METHODS

Strains: From eleven sequenced *Adh* alleles (KREITMAN 1983), we chose one Adh-fast and one Adh-slow allele representing the two electrophoretic ancestries at this locus (see Figure 1). The Japanese Adh-fast allele (Ja-f) has most of the characteristics of the Adh-fast ancestry plus six unique nucleotide changes. The African Adh-slow allele (Af-s) has the fewest number of differences compared to the consensus sequence for six slow alleles.

Subcloning and sequencing strategy: A 4.5-kb *Sal*I-*Sal*I fragment immediately upstream from the *Adh* locus was isolated from the Ja-f and Af-s *Adh* recombinant phage strains and subcloned into pUC8. Random shotgun clones were prepared from these fragments in M13mp8 (MESSING 1983) and were sequenced by the dideoxyribonucleo-

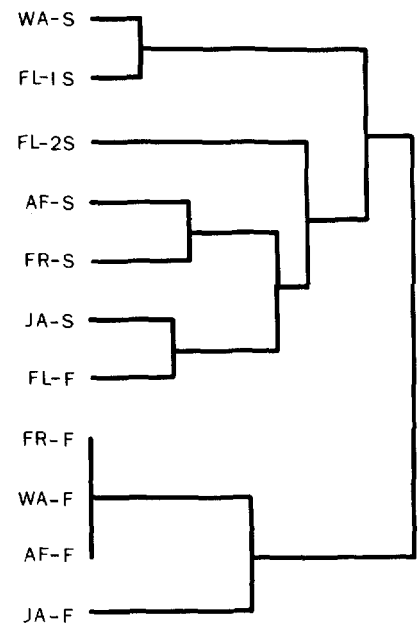

FIGURE 1.—Unweighted-Pair-Group (UPG) phenogram for eleven *Adh* alleles based on 43 sequence differences (from KREITMAN 1983; STEPHENS and NEI 1985). Ja: Japan; Af: Africa; Wa: Seattle, Washington; Fl: Southern Florida; Fr: France.

tide chain termination method (SANGER, NICKLEN and COULSON 1977; BIGGINS, GIBSON and HONG 1983). Complete sequences were determined from both strands.

Four-cutter filter hybridization analysis: Of the 87 isochromosomal second chromosome lines described in KREITMAN and AGUADÉ (1986), 81 were used to estimate levels of variation in the *Adh* 5′ flanking region using four-cutter analysis. Twenty-three lines were established from a Putah Creek (Davis, California) collection and 58 lines were established from a Raleigh Farmer's Market (Raleigh, North Carolina) collection. The procedure for establishing the isochromosomal lines, preparing DNA and preparing filters containing restriction digestions are described in KREITMAN and AGUADÉ (1986). The 4.5-kb *Sal*I-*Sal*I region upstream to *Adh* was probed using a gel-purified *Sal*I-*Sal*I fragment as described in KREITMAN and AGUADÉ (1986).

DNA sequences were manually aligned as described in COYNE and KREITMAN (1986) to minimize the total number of differences between sequences, counting both insertions/deletions and base changes as single differences.

RESULTS

Variation in two *D. melanogaster Adh* alleles. Figure 2 shows a restriction map of the *Adh* region in *D. melanogaster*, including the positions of the *Adh* structural locus and two other putative but otherwise uncharacterized open-reading-frames (ORFs). Two *D. melanogaster* sequences are presented in Figure 3 corresponding to a fragment beginning at a *Sal*I site 4.7 kb upstream from the *Adh* locus and ending 58 base pairs (bp) upstream from the transcription initiation site of the adult *Adh* mRNA. Discounting insertions and deletions, there are 4511 bp of aligned sequence. We have detected an ORF on the opposite strand to *Adh* starting with a methionine codon at position 436 and

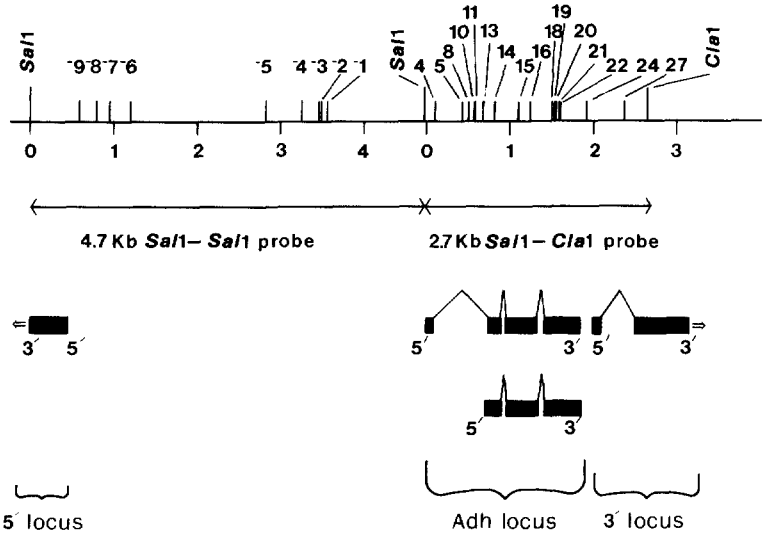

FIGURE 2.—Map of the *Adh* region in *D. melanogaster*. The 4.7-kb 5′ flanking *SalI-SalI* fragment (see Figure 3) and the 2.7-kb *SalI-ClaI* (*Adh* locus) fragment are numbered separately. Polymorphic restriction sites (see Table 2) are shown above the line. The *Adh* transcript and the putative 5′ and 3′ transcripts (see text) are shown below the line.

extending leftward away from the *Adh* locus. The ORF consists of 146 amino acids before reaching the end of the DNA sequence.

Analysis of both base composition and codon usage across the whole region (data not shown) suggests the presence of a coding region corresponding to the ORF, although we have not obtained conclusive direct evidence for an appropriate mRNA. Nevertheless, the first 500 bp of the *SalI* fragment (containing the ORF) have been excluded from the subsequent analysis of levels of variation flanking the *Adh* locus. No other large ORFs were identified in the *Adh* 5′ flanking region.

Comparison of the two sequences reveals 26 nucleotide and seven length differences in a total of 4713 bp of 5′ sequence. (Stretches of DNA that require multiple insertions/deletions and/or base changes in order to make an alignment are considered as single insertion/deletion events. Nucleotide changes embedded or immediately adjacent to such insertions/deletions are not considered to represent independent nucleotide changes.) These two sequences differ, in addition, at three nucleotide sites in the 58-bp region connecting the 3′ end of the sequence in Figure 3 to the 5′ start of the *Adh* locus (see KREITMAN 1983). There are no differences between the two sequences in the putative 5′ gene.

The distribution of nucleotide differences between the two alleles for the 5′ flanking region, the *Adh* locus and a 0.8-kb 3′ flanking region is plotted in Figure 4A as a histogram showing the number of nucleotide differences per 100 bp. A summary of these data is given in Table 1. Two statistical approaches were used to test for heterogeneity in levels of variation, both of which failed to reveal significant differences. The first evaluates whether there

is any heterogeneity in numbers of differences per 100 bp, *e.g.*, whether, at this scale, the number of differences are nonrandomly distributed. To do this, a goodness-of-fit test (SOKAL and ROLF 1969) between the observed frequencies of the classes 0, 1, 2, 3 and 4 + 5 + 6 and expected frequencies calculated from a Poisson distribution with an expected mean equal to the observed mean (0.68) reveals no significant deviation from Poisson ($G = 8.5$, $P > 0.1$).

The summary values given in Table 1 were used to evaluate whether there is heterogeneity in levels of variation in the 4-kb 5′ flanking region, the *Adh* locus and 0.8-kb 3′ flanking region. As indicated in the table legend, there is no significant heterogeneity ($G = 4.7$, $P > 0.1$).

The lack of statistical evidence of heterogeneity across the three regions, defined either functionally or physically, is surprising, since one of the regions contains the *Adh* structural locus. Overall, the frequency of polymorphism, as estimated by heterozygosity per nucleotide site (H), is actually slightly higher for the structural locus ($H = 0.01$) than for the 5′ flanking region ($H = 0.007$). In addition, nine of the 18 nucleotide differences in the *Adh* locus fall within the three translated portions of the gene, eight of which are silent. Since only 25% of all possible nucleotide changes in the coding region are silent (KREITMAN 1983), there are approximately 0.25 × 765 bp = 192 "effectively" silent sites. Therefore, the frequency of "effectively" silent polymorphism in the coding regions is 8/192 = 0.04, a value that is 5.7 times higher than the 5′ flanking region estimate.

In order to evaluate whether this pattern is simply an artifact of the particular evolutionary histories of the pair of alleles analyzed above, we investigated four-cutter restriction polymorphism in the same 7.2-kb region in 58 isochromosomal lines of *D. melanogaster* established from a 1983 Raleigh Farmer's Market (Raleigh, North Carolina) collection and in 23 isochromosomal lines established from a 1983 Putah Creek (Davis, California) collection (KREITMAN and AGUADÉ 1986). The restriction polymorphism analysis for a 2.7-kb *Sal*I-*Cla*I region containing the *Adh* locus (see Figure 2) is presented in KREITMAN AND AGUADÉ (1986). A summary of these data, as well as a summary of restriction site polymorphism in the 4-kb 5′ flanking region are given in Table 2. Ten restriction enzymes were used in the 2.7-kb *Sal*I-*Cla*I region encompassing the *Adh* structural locus. Six enzymes were used to survey restriction polymorphism in the 5′ flanking region: *Taq*I, *Alu*I, *Sau*3A, *Hae*III and *Dde*I + *Bam*HI.

Table 3 contains a summary of a 4-kb polymorphism based on the restriction data for a 6.7-kb region including 5′ flanking region, *Adh* locus and 3′ flanking region. There is no significant difference in the level of polymorphism in these three regions ($P > 0.1$). Again, there is actually a slightly higher level of polymorphism in the structural locus ($H = 0.006$) compared to the 5′ flanking region ($H = 0.004$). Therefore, the populational comparison yields the same results as the sequence comparison. There is no evidence, then, that the two sequenced alleles are unrepresentative of a random population sample. Both lines of statistical evidence indicate similar levels of heterozygosity for the *Adh* locus and its flanking regions.

```
         *         *         *         *         *        60
CATCCTCGCCCGTTTCCACGCCGTCGTCCTCCTCATCATCGGCGAGAGCTGATTGCGTGG
         *         *         *         *         *       120
TGGTCAGAGGCGAACCAGCGGTCTTCGTGGAGCTGGGACCCAGATCAAGGCTGCTCAACA
         *         *         *         *         *       180
GATTGCCTGCCGACTGGGAAGACGTTAGGGTGTCCTTGTGATAGGAGCTGTGCCGATTGC
         *         *         *         *         *       240
CCAGCTTAGTGGATAGTGTTAGGTCGCCGTTGCTCGTTGGGCGTAGACTGCCCACCACCT
         *         *         ●         *         *       300
GACCACCGGGCAGGGTGGCGCTTCTCTTGTGGCGACCCTTCGACTTGGGAAAGGCAGCCA
         *         *         *         *         *       360
GGATGTTGAGCCACCACTGGGATTCCTCTGAACTGGTGCCCTTCACAAAGGTCACGCGCT
         *         *         *         *         *       420
CGGGAGCGGTTATGGCGATGGAGTTGGGGTGACCTGTCACCTCCACGGCGCTGGTAACCT
         *         *         *         *         *       480
CCAGCACTTTGGTCATATCAACGCACGCCTGCGGTATGGTTTCGGGCTATAGAAAATATA
         *         *         *         *         *       540
TGTAAATTAAAGAGTAAACAAGTTGTATTTTAAGATTTTAATTAGGAGAATTAATTAATC
         *         *         *         *         *       600
GGTAATCACATGAACTCGGCCTATCGCGTAATAATATACATTTTTTAATTTAATGACTAA
        A
         *         *         *         *         *       660
TAAATAATATAAAATCTAATTAATAGTTCAGTAAGTTAGTAAAAGTAAATCAATCTGGTG
         *         *         *         *         *       720
GTAATTTAAGAAGCCACTTTAATTCTTCCACTTCATAAATAATCGGCTGGTTAAGGAAAG
 C       A                   T
         *         *         *         *         *       780
GTACATTTATTGTTGTTTTTACCGCCAGCACACTTATTGGTTCTACCGATAACGTCCAGC
         *         *         *         *         *       840
GATGCTATAATACCATAATTAGAAGCTCTTTTGGGATTGTATAATATTTTATGAGCTTTG
A
         *         *         *         *         *       900
TTATCTTATAAATTCAGACCACCCCATAACAGAGTTTTATTATCTTTTTATTTTTTTGTT
                                ↑AA
         *         *         *         *         *       960
ATCACTGGAGAACCAACGAGACGGTATTTAAACAGAAAAATACAATTATGCCTATGGATT
                                  ↑A     C
         *         *         *         *         *      1020
GATTAGCTATTACAAACTCAAAAATTCGATTTAATTTTATTATTAGCTATAAAAATGGAA
                                T
         *         *         *         *         *      1080
ATGGTTTAAAATATGTTCAAATGAATTACTTACATAATCATCAACCGAGTACGTCAGCTC
         *         *         *         *         *      1140
GCCATCATCATAGAGAACAAACCATCTTCTTTGCCAGCGCTACAATTGAAAAAGAAGACA
         *         *         *         *         *      1200
AATTTTATTAAAAATATTAAACTATTCTCAAGTTTATATTATTTGATCTTTACTAAGTCT
                                  C
         *         *         *         *         *      1260
AAGTCTGTGGCTATCAGTCGATGAGAGTGATCAACTCTAAAACAAATTTACATTGTCGCT
         *         *         *         *         *      1320
TGCAATTTGCAACATGAAAGGTGGGACGAGAAATGGTGAGGAAAGACAAGATCGGATGTA
```

Figure 3 (see legend on page 101).

```
        *         *         *         *         *      1380
AATAATGTTCAACGCCCCCGACAGAAATCATAATTCCTTTATAATTCGTTCTTTCATAAA
        *         *         *         *         *      1440
TTTTCAGGCGTTGTCATTCCCATGAAAGGCAACCAAGCCCCAAACGCCTTCGCCTTTGCA
        *         *         *         *         *      1500
TTGGCACTGATTCGCTGTGGATCTGGATCTCTATCTGTATCTGCATCTGTATCTGTATCT
        *         *         *         *         *      1560
GAATATGAATCTGAATCGGAATCTGGATCTGATTCTCATTGTTATTGTTGGTTGCCAGAA
        *         *         *         *         *      1620
TCATAACAAACGTGCAACAGCCACAAGGGTATAGGACTCAACGTGTGTCTGATATTTATG
        *         *         *         *         *      1680
CAAATTGTTAAAAGTCAAAGCAAATTAAGCTCAACCTTCAGCGAAGATGACGTTGAATTC
        *         *         *         *         *      1740
TGTTGCCCTATTGCGCTGTAAGTTGCTAGTTGCAAGTTGCAAGTTGCACCTTTCTGCAGT
        C         C
        *         *         *         *         *      1800
TGATTTCTCCTCATCCACCTATGCAGTCAGGTGAGAGGGAGTGAGTGCGAGTGGAGTGCT
        *         *         *         *         *      1860
GAGGTGTGTCAAGCGAATTATTTATAAGGCCTAGAAGAAGGCAGCTCGCACGCGAATAAT
        *         *         *         *         *      1920
CAAGACTCAGCACCAATTTTTAGTTTATGGTCTAGTTCTTTATAGGTTTTGTACTTCTTT
                                                     C
        *         *         *         *         *      1980
TTTTTGCGTTGGCTATTTTGCGATTGAATTCATAAATATGGAATCAAATCTATAGAGTGG
        *         *         *         *         *      2040
AGAGTGGAACTAACGAGGTGAGAGGTAACAATATAGTTTTTCGGCAATCAGAAGCAACAA
                                          G
        *         *         *         *         *      2100
ACAAATATCTGCAATAACTCGTTGAATTCGAAACAAAATTAACTGCATTTATACTAAATA
        *         *         *         *         *      2160
TATAATTGCTATAGGATGAGTTAGCCGTCTTGCGGTTTCCCAAACCCCAAAAGCAAAGTC
        *         *         *         *         *      2220
AAGCGTGTAGGAAACCTGATCAGATCGCGGGAAAGATTCTCTGCACTCAATTACGTCAAA
        *         *         *         *         *      2280
CCAGGTTGATTTCCTCCTTTTCGCTGTCGAGAGATTGGCAAATGGGTCAAATGGGTCAGG
        *         *         *         *         *      2340
CAGTGGAATAGTAAATTAGATTATGTTTGCATCGAGATGCAATGCAAGCCGCGCCCCAAA
        *         *         *         *         *      2400
TAAATGGAACGTGCGCTAGTAGGGTTCCCCCTTGCCCCTGGTAACCCTTCCTTTCACCAC
        *         *         *         *         *      2460
CCGTTTTCCCGCTTTTCCGCTCCCAAACACTAGAGGTAAGCTGCTTAGACCCCGGCGTTT
        *         *         *         *         *      2520
AGAAGCCCCAGTTTCGTTTCACTAGGCAGACACACTCGCAGCGGGAAGACAATGCCATCG
                                                       ---
        *         *         *         *         *      2580
CCACCGCCACCGACTTAATCAGCCCGCGAAACGACATCTCAATGCTGGCGAGCGTGTACC
------------------------------------------------------------
        *         *         *         *         *      2640
TACATATGGACATGGGCGTGCGTTGGTGCGGGAGCTGGTGTAAATCGGTTTTGGCAGGTA
------------------------------------------------------------
```

Figure 3—*Continued.*

```
        *         *         *         *         *      2700
CGCCGCTGGCGTCATTACCCCCCAGAGGTTGAATGTCACCGGCGGCATGACTTGGGGGCC
---------------------------------------------------------
        *         *         *         *         *      2760
AAGCCGATAAGGCGCACACTGTCCACTGCACGGTGTACACTGATAAAAATATATATCAAG
-----↑AT
        *         *         *         *         *      2820
ACCAAATACTGTTAAAGATAATTGATGCGTAAAGGAAATACACTTGCAAGTTAAAATGTT
        T                   T
        *         *         *         *         *      2880
TTCACCTTAATGTGTTTTTCTTTTAATACTCTATTAACTAATATAAATTATCACCAAAAC
                          ↑A
        *         *         *         *         *      2940
AAAACATTAATTTGGGAAATGTTATCACCAAAAGCTTTTGCCACTATAGAAAATACAGAT
        *         *         *         *         *      3000
AAATCTAAAAATAAATTCCTTTGACGTATGCACGAAATAAGATAAACAAATTTGATTTTA
                        C
        *         *         *         *         *      3060
TTTTCTTATTTAAACAATTCATTTTATTTGCATGCATGCGTATGCCAATCTATTTTGTTC
            C                         ----------↑T-
        *         *         *         *         *      3120
AGTGTACCTAATAAAAACGATTTCGTTTGCCCCAAGTAGTAAGAAGATGTTAGGCACGTC
                                      C
        *         *         *         *         *      3180
TGCTGATAAGGAAAACTGTAGCCCCAGACTAGGCCAGACCATATTAAATTAACGTCTGGA
        *         *         *         *         *      3240
GGCGCGAACAGTCATACGATTTTTTTTTTATATTACTTCACGGTCAGTTGCCAAGGCAGG
                              -         G
        *         *         *         *         *      3300
AGAGCAACCCGTTCGATTAGTGGGTCAATTTGGAAAATGAGTTATTGACTCTGGGAAATT
        *         *         *         *         *      3360
GTTGAGCTGAAAATTTAATCAGAGCCCGAAAATTTCCAATCATGCATTCCCCAAGTGACC
                    G
        *         *         *         *         *      3420
ATATATGGATTAGTGATAACGCTCGATGCGACCCCCAAAGATTATCAAAAATATTTAATA
        *         *         *         *         *      3480
TGAATATATGAAAAAAAGATTTAACTTTTATGAATTCTTAAGCGTCCCCAAAGCTTCGGG
A                                     C         T
        *         *         *         *         *      3540
AGAACTGGGCCATATATGACCCGAAATACATGTTTATACTTTAGCAAATGTATTTTCCAA
        *         *         *         *         *      3600
TTAGGTGATAGAACTTGTGTGCACACACACATATAGTTCTATATCAACAAACAGGTTTAA
        *         *         *         *         *      3660
GTTTTATGCAAATTGAAAGCTTATTTCTTCCGCATGCTTATCTCTTTCCTTCTCATCATT
        *         *         *         *         *      3720
TGTATGCAAAAAATACATATGAATTTGCAGTAGCCTCCTCCCACATCATATTTAACGCCC
        *         *         *         *         *      3780
TATATTCAAAATTTGCTCAAGAAAATATTTGAACCAAATTGATTTTTAGTCAATTAGTTT
        *         *         *         *         *      3840
TTAAGTAATTAAGTGGAGTAAACATATACAATTTTATTCTTACCAAACACATATACTCAT
```

Figure 3—*Continued.*

```
        *         *         *         *         *      3900
ATATTTTGAATAAATAAATAAACAAATATATATAAAATCTACGAAATTGGCAAACAAATT
        *         *         *         *         *      3960
TTAAAGCATTATAGTATTGCCGATTTAATTAATATAATTAAATAATATGTACATGTATTA
        *         *         *         *         *      4020
ATCTTGTGTGCGAGCATGGGTTAAATCTAGCTGCATTCGAAACCGCTACTCTGGCTCGGC
        *         *         *         *         *      4080
CACAAAGTGGGCTTGGTCGCTGTTGCGGACAAGTGAGATTGCTAATGAGCTGCTTTTAGG
        *         *         *         *         *      4140
GGGCGTGTTGTGCTTGCTTTCCAACTTTTCTAGATTGATTCTACGCTGCCTCCAGCAGCC
        *         *         *         *         *      4200
ACCCCTCCCATCCCCATCCCCATCACCATCCAGTCCCGTTGGCTCCCAGTCACAGTATTA
          A   T
        *         *         *         *         *      4260
CACGTATGCAAATTAAGCCGAAGTTCAATTGCGACCGCAGCAACAACACGATCTTTCTAC
        *         *         *         *         *      4320
ACTTCTCCTTGCTATGCTTGACATTCACAAGGTCAAAGCTCTTAATATTCTGGCTCGTGG
        *         *         *         *         *      4380
CCCTACACTGTAAGAAATTACTATAGAAATAACGGTACACGGAATAAGATATTTTTTTTA
        *         *         *         *         •      4440
GTCCATATGCTTTTAACAAATGTGTTTTGAGTTTATGTTATATTATTGTTAGAAAACCGG
                            A
        *         *         *         *         *      4500
TGTTTTTTTTTAAATCGGTTAAAAAATTACTACGAGAGAAAAATACAAATTTTGTAAATA
        -A
        *         *         *         *         *      4560
AGATTGACTCTTTTTCGATTTTGGAATATTTTCATTCATTTTATGTTTTTACGTTTTCAC
        *         *         *         *         *      4620
TTATTTGTTTCTCAGTGCACTTTCTGGTGTTCCATTTTCTATTGGGCTCTTTACCCCGCA
        *         *         *         *         *      4680
TTTGTTTGCAGATCACTTGCTTGCGCATTTTTATTGCATTTTACATATTACACATTATTT
        *         *       4713
GAACGCCGCTGCTGCTGCATCCGTCGACGTCGA...3*
```

FIGURE 3.—DNA sequence of the 4.7-kb *Sal*I-*Sal*I *Adh* 5′ flanking region for an Adh-s allele (Af-s). Nucleotide differences between this allele and an Adh-f allele (Ja-f) are shown below the sequence. –, deleted position; ↑, sequence inserted at next position.

Interspecific comparison of *Adh* alleles: Selective constraint on amino acid replacement sites are now well documented at the *Adh* locus in Drosophila (KREITMAN 1983; BODMER and ASHBURNER 1984; COHN 1985; SHAEFFER 1985). KREITMAN (1983), for example, estimates that the silent coding positions are four times more polymorphic than a site chosen at random from the entire locus. The fact that the 5′ flanking region exhibits roughly the same level of polymorphism as the entire *Adh* locus raises the possibility of some considerable selective constraint in the 5′ noncoding region as well.

One way to investigate this possibility is to test whether interspecific comparisons also reveal reduced variation in the flanking DNA. To do this, we have compared levels of sequence divergence in the *Adh* locus and its flanking regions in two species, *D. melanogaster* and its sibling species, *D. simulans*. A high level of divergence in the 5′ flanking region relative to the structural

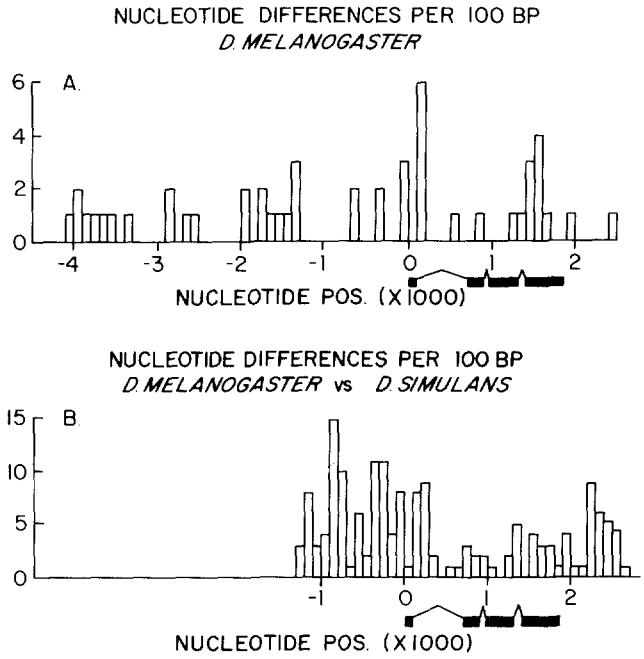

FIGURE 4.—Histogram of nucleotide differences per 100 bp of aligned sequences. A, Af-s *vs.* Ja-f; *B*, *D. melanogaster* vs. *D. simulans* (COHN 1985). Summaries are given in Tables 1 and 4.

TABLE 1

Distribution of nucleotide variation based on sequence differences in two *D. melanogaster* alleles (Ja-f and Af-s)

		Adh locus				
			Introns			
	5′ flanking	Nontrans-lated	1	2 + 3	Coding	3′ flanking
Base pairs	4000	335	620	135	765	800
No. of polymorphic sites						
Observed	29	1	7	1	9	2
Expected	29.4	2.5	4.6	1.0	5.6	5.9
Heterozygosity per nucleotide	0.007		0.01			0.003

Goodness of fit (SOKAL and ROHLF 1969) for observed and expected number of nucleotide differences in the 5′ flanking, *Adh* locus and 3′ flanking regions: $G = 4.7$; $P > 0.1$.

locus, often evident in interspecific sequence comparisons, would be considered *prima facia* evidence for lack of selective constraint in the flanking DNA.

Summary of a sequence comparison between the Af-s allele of *D. melanogaster* and a *D. simulans* allele sequenced by VIVIAN COHN (COHN 1985) is given in Table 4, and the distribution of nucleotide differences per 100 bp is given in Figure 4B. In contrast to the two previous within-species comparisons, the 1325-bp 5′ flanking sequence is twice as divergent as the *Adh* locus (6.0 *vs.*

TABLE 2

Polymorphic four-cutter restriction site frequencies for two *D. melanogaster* populations (Pu and Ra)

Site[a]	Enzyme	Position	Allele	Pu	Ra	Heterozygosity
−9	*Hae*III	558–561	+	23	56	0.048
			−	0	2	
−8	*Alu*I	834–837	+	22	56	0.024
			−	1	0	
−7	*Taq*I	989	+	6	32	0.498
			−	17	26	
−6	*Dde*I	1208 or 1781	+	0	1	0.024
			−	23	57	
−5	*Dde*I	2830	+	0	1	0.024
			−	23	57	
−4	*Taq*I	3253–3256	+	18	52	0.235
			−	5	6	
−3	*Dde*I	3458	+	9	28	0.384
			−	14	20	
−2	*Alu*I	3472–3475	+	22	54	0.116
			−	1	4	
−1	*Alu*I	3526 or 3577	+	2	2	0.094
			−	21	56	
4	*Ban*I	107	+	7	29	0.485
			−	20	31	
5	*Alu*I	423	+	3	2	0.108
			−	24	58	
8	*Msp*I	502–505	+	24	58	0.108
			−	3	2	
10	*Hha*I	571–574	+	25	55	0.148
			−	2	5	
11	*Msp*I	586	+	26	59	0.045
			−	1	1	
13	*Hae*III	687–690	+	25	56	0.128
			−	2	4	
14	*Hae*III	816	+	15	25	0.497
			−	12	35	
15	*Alu*I	1068	+	27	59	0.023
			−	0	1	
16	*Msp*I	1235	+	26	58	0.067
			−	1	2	
18	*Dde*I	1518	+	20	48	0.341
			−	7	12	
19	*Dde*I	1527	+	7	12	0.341
			−	20	48	
20	*Dde*I	1551	+	1	0	0.023
			−	26	60	
21	*Hae*III	1563–1566	+	26	58	0.067
			−	1	2	
22	*Alu*I	1596	+	23	51	0.254
			−	4	9	
24	*Hae*III	1925	+	27	53	0.148
			−	0	7	
27	*Taq*I	2348–2351	+	27	54	0.128
			−	0	6	

[a] Position numbering based on Figure 3 (this text) for 5′ flanking region and KREITMAN (1983, figure 3) for *Adh* and 3′ flanking regions.

TABLE 3

Distribution of nucleotide variation based on four-cutter restriction differences in a sample of 81 alleles from two *D. melanogaster* populations (Pu + Ra)

	Adh locus					
			Introns			
	5′ flanking	Nontrans-lated	1	2 + 3	Coding	3′ flanking
Base pairs	4213	335	620	135	765	811
Site equivalents[a]	414	42	102	16	251	129
No. of polymorphic sites						
Observed	9	0	6	0	8	2
Expected	10.8	1.1	2.7	0.4	6.6	3.4
Heterozygosity per nucleotide[b]	0.004		0.006			0.002

Goodness of fit for observed and expected number of polymorphic sites in the 5′ flanking, *Adh* locus and 3′ flanking regions: $G = 1.9$, $P > 0.1$.

[a] Site equivalents are calculated as the product of the length of a region (bp) times the fraction of all possible changes in a consensus sequence that would be detected as restriction site changes (see KREITMAN and AGUADÉ 1986).

[b] Heterozygosity per nucleotide site is calculated as the sum of the heterozygosities for segregating sites (Table 2) divided by the site equivalents in the region.

TABLE 4

Distribution of nucleotide divergence between *D. melanogaster* (Af-s) and *D. simulans* (COHN 1985)

	Adh locus					
			Introns			
	5′ flanking	Nontrans-lated	1	2 + 3	Coding	3′ flanking
Base pairs	1325	335	620	135	765	800
No. of polymorphic sites						
Observed	86	9	21	6	12	31
Expected	54.9	13.9	25.7	5.6	31.7	33.1
Divergence	0.06		0.03			0.04

Goodness of fit for observed and expected number of nucleotide differences in the 5′ flanking, *Adh* locus and 3′ flanking regions: $G = 27.9$, $P < 0.001$.

3.0% divergence, respectively). There is a statistically significant heterogeneity in the distribution of substitutions in the structural locus and two flanking regions ($P < 0.001$). Therefore, whereas the two *D. melanogaster* alleles show a similar level of divergence in the 5′ flanking and structural regions, the interspecific comparison shows a clear pattern of conservation in the structural locus relative to the 5′ flanking sequence. In fact, there is no significant difference between divergence levels for "effectively" silent coding sites and the 1325-bp 5′ flanking region ($P > 0.5$). We reject, therefore, the hypothesis that

the 5′ flanking region is differentially constrained relative to effectively silent coding sites.

DISCUSSION

The analysis presented above investigates evolutionary constraints in the *Adh* locus and its flanking DNA by joint analysis of intraspecific and interspecific nucleotide data. The original observation suggesting reduced evolutionary rates in the noncoding DNA flanking the *Adh* locus was KREITMAN'S sequence study of eleven *Adh* alleles and their 3′ flanking regions (KREITMAN 1983). In that study, an 846-bp 3′ flanking region showed approximately a tenfold reduction in the frequency of segregating sites compared to the two small introns and "effectively" silent sites in the exons. This led him to hypothesize either a lower mutation rate in the flanking region or some unknown selective constraint.

The latter hypothesis appears to be correct: the conserved region is now believed to contain, in large part, the coding region of another gene (COHN 1985; SHAEFFER 1985). Nevertheless, the frequency of segregating sites in the *Adh* coding region, the most highly conserved part of the transcriptional unit, is three times higher than in the 3′ flanking region containing the putative 3′ gene (1.9 *vs.* 0.6% divergence, respectively). Therefore, a problem still remains of explaining the apparent excess of polymorphism at silent positions in *Adh*.

Because of the uncertainty about the structure of the putative 3′ locus we have, instead, concentrated on the 5′ flanking region. To evaluate whether the lower level of polymorphism oberved in a 4-kb flanking sequence could be evidence for functional constraints in this region, we compared levels of sequence divergence between *D. melanogaster* and its sibling species, *D. simulans*. This analysis reveals a twofold higher nucleotide divergence in the *Adh* 5′ flanking region than in the structural locus. In contrast to the within-species results, there is essentially no difference in nucleotide divergence between the 5′ flanking region and the "effectively" silent sites in the *Adh* coding region.

If nucleotide substitutions in the two regions are selectively unconstrained (or are similarly constrained), this comparison would be a direct test of the equivalence of mutation rates in the two regions. This is because, as KIMURA has shown (KIMURA 1968), the substitution rate for effectively neutral mutations depends only on the mutation rate. Without prior knowledge about patterns of constraint in regions of interest, sequence comparisons of different regions of DNA essentially test the compound hypothesis for equality of mutation rates and equivalence of selective constraints. The lack of a statistically significant difference in divergence between the 5′ flanking region and the "effectively" silent sites in the *Adh* coding region suggests that there is no differential selective and/or mutational pressures.

The possibility remains, however, that some or most of the 5′ flanking DNA not considered in the interspecific comparison contains one or more "hidden" structural loci, leading to the observed low level of polymorphism in this region in *D. melanogaster*. In fact, we have identified a large open-reading-frame approximately 4 kb upstream from *Adh*, and for this reason have excluded this

region from the analysis of the 5′ flanking region. There are two lines of evidence mitigating against another structural gene being "hidden" in the approximately 1.8-kb region between the 5′ ORF and the rapidly evolving proximal DNA identified by interspecific sequence comparison.

First, the entire region is relatively rich in adenosine and thymine (70%), with frequent homonucleotide runs, a common occurrence in DNA flanking Drosophila genes. In accord with this observation is the paucity of ORFs in this region (the largest ORF is 34 codons) and a lack of evidence for a coding region, based on Pustell's codon usage-biased search routine (distributed by IBI Corporation). (This program successfully identifies the *Adh* locus exons and also identifies the 5′ ORF.)

Second, we have identified seven insertion/deletion differences between the two sequenced alleles, all of which occur between the 5′ ORF and the *Adh* locus. We have also identified 15 insertions/deletions in this 4.5-kb flanking DNA in the 81 lines scored by four-cutter restriction analysis (data not shown). On the basis of these two lines of evidence, we tentatively conclude that there are no coding regions between the ORF and *Adh*.

Before considering some possible explanations for the relatively higher level of polymorphism in the *Adh* structural locus compared to the 5′ flanking region in *D. melanogaster*, we first note that this finding is based on two independent observations: a sequence comparison of the complete 7.2-kb region in two alleles and a four-cutter restriction analysis of 81 wild isochromosomal lines. The two studies give remarkably similar results. Although the heterozygosity per nucleotide is higher in the sequence comparison than in the population analysis for all three regions (this is expected since the two alleles were chosen to be different), the relative estimates are virtually identical. For example, both studies show the *Adh* structural locus to be the most polymorphic and the 3′ flanking region to be the least polymorphic. Furthermore, within the structural locus, both studies identify the same pattern of polymorphism in the introns, exons and nontranslated regions.

One possible criticism of the statistical tests employed here is that the sites under consideration are tightly linked and are unlikely to be highly recombining either within or between adjacent regions. In such a case, the evolutionary fate of independent but tightly linked mutations would not be independent of one another; therefore, the distribution of segregating sites would not be Poisson. Because the evolutionary histories of tightly linked sites are correlated, the expected variance in the number of segregating sites will actually be larger than Poisson (WATTERSON 1975). We note, then, that the significance levels given here are overestimates: the actual significance value under a more realistic model is expected to be lower. However, even using an infinite allele model with no recombination within regions but free recombination between regions, which is expected to overestimate the actual variance of the number of segregating sites, the same data are still significantly different from the neutral expectation (R. HUDSON, personal communication).

Therefore, the different levels of silent polymorphism at the *Adh* locus and the 5′ flanking DNA in *D. melanogaster* appear to be too large to be reconciled

TABLE 5

Distribution of nucleotide variation within and between species: *Adh* locus *vs.* 3′ flanking region

Comparison	*Adh* locus	3′ flanking
Within species[a]		
Observed (length)	35 (1855)	5 (767)
Expected	28.1	11.9
Between species[b]		
Observed (length)	48 (1855)	31 (800)
Expected	54.8	24.2

Test of independence: $G = 9.3$; $P < 0.005$.
[a] *D.melanogaster* (from KREITMAN 1983).
[b] *D. melanogaster* vs. *D. simulans* (Table 4).

by either of the standard neutral models (*e.g.*, infinite site and infinite allele model). This leads us to consider alternative explanations for the observed pattern of variation.

The first explanation to consider is the possibility that the 5′ flanking region in *D. melanogaster* has gained additional functional constraints leading to a reduced level of variation. Such a loss of variation could also have resulted from a recent fixation of a selectively favored mutation in the 5′ noncoding region. However, there are two reasons to think that this is not the case. First, there is also a reduction in the level of variation in the region 3′ to the *Adh* locus (see DISCUSSION above). Table 5 gives within-species heterozygosity estimates and between-species divergence estimates for the *Adh* locus and for an 800-bp region 3′ to the locus based on KREITMAN'S sequence comparison of 11 *D. melanogaster* alleles (KREITMAN 1983) and on COHN'S sequence of a *D. simulans* allele (COHN 1985). There is a significant difference in the within- *vs.* between-species distributions ($P < 0.005$). Again, the difference can be explained either by an excess of silent polymorphism in the *Adh* locus or a reduction in polymorphism in the 3′ region. Thus, there is a reduction of polymorphism both 5′ and 3′ to the *Adh* locus in *D. melanogaster* relative to the distribution of differences between the species.

Second, the estimate of heterozygosity for the *Adh* locus is consistently higher than other estimates of heterozygosity in *D. melanogaster*. For example, two different studies of restriction polymorphism in a 12-kb region encompassing the *Adh* locus (LANGLEY, MONTGOMERY and QUATTLEBAUM 1982; AQUADRO *et al.*, unpublished results) have reported an average heterozygosity per nucleotide of 0.006 for the region, whereas our population estimates are consistently around 0.02 for silent nucleotide positions and introns. This suggests that the larger region surrounding the *Adh* locus is less polymorphic than are silent positions within the locus.

Given that there is an excess of variation in the *Adh* locus, rather than a reduction of variation in the flanking regions, it is also possible to explain this observation as a recent loss of constraints within the locus in the *D. melanogaster* evolutionary lineage. A relaxation of constraints would then lead to an

increase in polymorphism within the species. However, depending on how soon after the species split the relaxation occurred, some effect would be expected on the distribution of substitutions in the between-species comparison as well. Thus, this hypothesis places some limitations on when in the evolutionary history of the species the loss of constraint could have taken place. In addition, since the two small introns as well as the silent sites in the coding regions (but not the replacement sites) show similarly high levels of polymorphism, relaxed constraints on both kinds of changes must be hypothesized.

A third explanation for the excess polymorphism in the *Adh* coding regions involves a selective maintainence of the *Adh* protein polymorphism. STROBECK (1980, 1983) has shown that the expected heterozygosity for a neutral locus linked to a balanced polymorphism increases with decreasing recombination. This excess heterozygosity results from the evolutionary correlation of tightly linked polymorphisms.

In support of this explanation is the observation that exon 4, which contains the amino acid polymorphism distinguishing the two allozymes, also has the highest level of silent polymorphism. For example, in the sequence comparison of 11 *Adh* alleles, 3.5% of all sites are polymorphic in exon 4, compared to 1.0% polymorphism in exons 2 and 3 (KREITMAN 1983). Similarly, six of eight coding region polymorphisms identified in the restriction study of KREITMAN and AGUADÉ (1986) are in exon 4.

However, if "hitchhiking" between silent polymorphisms and a hypothesized balanced polymorphism at the amino acid replacement site is responsible for the excess polymorphism around that site, then this excess should largely be segregating between, rather than within, the two allozymes. No excess polymorphism would be expected within an allozyme class. This is not the case, however. Recalculating heterozygosity from the restriction polymorphism data given in Table 3, considering only Adh-slow alleles, heterozygosity in the 5′ noncoding region decreases from 0.004 to 0.003 (sample size = 45), whereas heterozygosity in the *Adh* structural locus increases slightly from 0.006 to 0.008 (sample size = 53). Therefore, there is actually a slightly higher excess polymorphism in the structural region of the Adh-slow allele than there is for the population sample including both allozymes. This would suggest, then, that the excess polymorphism in the *Adh* structural locus, if it is a historical result of an association with a balanced polymorphism, would require a selectively balanced polymorphism other than the allozyme polymorphism. As such, the data presented here, while suggestive of one or more selectively maintained polymorphisms, cannot be taken as evidence supporting the two allozymes being a balanced polymorphism. Unfortunately, although we have tentatively ruled out the amino acid polymorphism as being the only site under balancing selection, it is not clear that this kind of statistical analysis will be able to resolve or delimit which nucleotide polymorphism(s) are under selection.

Selective events, such as those leading to adaptive change or balanced polymorphism, are recorded in DNA as changes in the location or amount of variation, but they may have short-lived evolutionary effects. This makes the study of variation within and between closely related species a necessity for

understanding the evolutionary dynamics of a gene. Most importantly, it provides a basis for identifying evolutionarily transient changes in the pattern of nucleotide differences between closely related alleles. The analysis presented here shows that information about the distribution of variation among alleles provides a framework for testing predictions of evolutionary models. In addition, this kind of analysis is useful for generating specific predictions about levels of variation at other loci and in other species. For example, there is an obvious need to obtain estimates of heterozygosity at other loci in *D. melanogaster*. Similarly, it would be useful to know something about the pattern of variation around the *Adh* locus in *D. simulans*. If the pattern of variation in *D. melanogaster* reflects selection in the recent evolutionary history of that species, then *D. simulans* would not be expected to show the same pattern of variation.

We thank C. LANGLEY, R. HUDSON, N. KAPLAN, J. C. STEPHENS, B. WALSH and W. QUATTLEBAUM for their comments and criticisms. This work was supported by National Institutes of Health grant GM29301 to R. C. Lewontin.

LITERATURE CITED

BIGGINS, M. D., T. J. GIBSON and G. F. HONG, 1983 Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80:** 3963–3965.

BODMER, M. and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding for *alcohol dehydrogenase* in sibling species of Drosophila. Nature **309:** 425–430.

COHN, V. H., 1985 Organization and evolution of the *alcohol dehydrogenase* gene in Drosophila. Ph.D. Thesis, University of Michigan, Ann Arbor.

COYNE, J. A. and M. KREITMAN, 1986 Evolutionary genetics of two sibling species of Drosophila. Evolution **40:** 673–693.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. Genetics **61:** 893–903.

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

KREITMAN, M. and M. AGUADÉ, 1986. Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by four-cutter hybridization. Proc. Natl. Acad. Sci. USA. **83:** 3562–3566.

LANGLEY, C. H., E. MONTGOMERY and W. F. QUATTLEBAUM, 1982 Restriction map variation in the *Adh* region of Drosophila. Proc. Natl. Acad. Sci. USA **79:** 5631–5635.

MESSING, J., 1983 New M13 vectors for cloning. Methods Enzymol. **101:** 20–79.

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

SOKAL, R. R. and F. J. ROLF, 1969 *Biometry*. W. H. Freeman, San Francisco.

SHAEFFER, S., 1985 Ph.D. Thesis, Department of Genetics, University of Georgia, Athens.

STEPHENS, J. C. and M. NEI, 1985 Phylogenetic analysis of polymorphic DNA sequences at the *Adh* locus in *Drosophila melanogaster* and its sibling species. J. Mol. Evol. **22:** 289–300.

STROBECK, C., 1980 Heterozygosity of a neutral locus linked to a self-incompatibility locus or a balanced lethal. Evolution **34:** 779–788.

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

WATTERSON, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Pop. Biol. **7:** 256–276.

Communicating editor: M. TURELLI