Estimates of Linkage Disequilibrium and the Recombination Parameter Determined From Segregating Nucleotide Sites in the Alcohol Dehydrogenase Region of *Drosophila pseudoobscura*

Stephen W. Schaeffer and Ellen L. Miller

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received January 13, 1993
Accepted for publication July 2, 1993

ABSTRACT

The alcohol dehydrogenase (*Adh*) region of *Drosophila pseudoobscura*, which includes the two genes *Adh* and *Adh-Dup*, was used to examine the pattern and organization of linkage disequilibrium among pairs of segregating nucleotide sites. A collection of 99 strains from the geographic range of *D. pseudoobscura* were nucleotide-sequenced with polymerase chain reaction-mediated techniques. All pairs of the 359 polymorphic sites in the 3.5-kb *Adh* region were tested for significant linkage disequilibrium with Fisher's exact test. Of the 74,278 pairwise comparisons of segregating sites, 127 were in significant linkage disequilibrium at the 5% level. The distribution of five linkage disequilibrium estimators D_{ij}, D^2, r_{ij}, r^2 and D'_{ij} were compared to theoretical distributions. The observed distributions of D_{ij}, D^2, r_{ij} and r^2 were consistent with the theoretical distribution given an infinite sites model. The observed distribution of D'_{ij} differed from the theoretical distribution because of an excess of values at -1 and 1. No spatial pattern was observed in the linkage disequilibrium pattern in the *Adh* region except for two clusters of sites nonrandomly associated in the adult intron and intron 2 of *Adh*. The magnitude of linkage disequilibrium decreases significantly as nucleotide distance increases, or a distance effect. *Adh-Dup* had a larger estimate of the recombination parameter, $4Nc$, than *Adh*, where N is the effective population size and c is the recombination rate. A comparison of the mutation and recombination parameters shows that 7–17 recombination events occur for each mutation event. The heterogeneous estimates of the recombination parameter and the inverse relationship between linkage disequilibrium and nucleotide distance are no longer significant when the two clusters of *Adh* intron sites are excluded from analyses. The most likely explanation for the two clusters of linkage disequilibria is epistatic selection between sites in the cluster to maintain pre-mRNA secondary structure.

The observation of significant linkage disequilibrium that is consistent between populations is a very sensitive detector of natural selection (LEWONTIN 1974).

LINKAGE disequilibria or nonrandom associations among genetic markers may result from forces other than selection such as low recombination rates, random genetic drift, or population subdivision (KIMURA and OHTA 1971; NEI 1987). LEWONTIN (1974) suggests that neutral forces fail to explain a consistent pattern of linkage disequilibrium in geographic populations unless the migration rate is sufficiently large to homogenize gene frequencies between populations. With extensive migration, population sizes would be too large for random genetic drift to maintain consistent patterns of nonrandom associations between linked loci.

Theoretical studies of the sampling distribution of linkage disequilibrium have focused on nonrandom associations between pairs of loci or sites to determine the importance of random genetic drift and selection as forces that generate linkage disequilibrium (HILL and ROBERTSON 1968; KARLIN and MACGREGOR 1968; SVED 1968; OHTA and KIMURA 1969a,b, 1971; SVED 1971; HILL 1975, 1977; STROBECK 1983; GOLDING 1984; HUDSON 1985). The expected value of linkage disequilibrium for a pair of selectively neutral loci in finite equilibrium populations is zero, although the variance of linkage disequilibrium can be quite large between populations (HILL and ROBERTSON 1968; OHTA and KIMURA 1969a). The alcohol dehydrogenase (*Adh*) region of *Drosophila pseudoobscura* is a model system to determine how much linkage disequilibrium exists in a selectively neutral region of the genome.

The *Adh* region is composed of two closely linked genes, *Adh* and *Adh-Dup* (SCHAEFFER and AQUADRO 1987), whose pattern and organization of nucleotide diversity failed to suggest the past action of positive Darwinian selection (SCHAEFFER and MILLER 1992b). In addition, evidence from nucleotide data indicate that geographic populations of *D. pseudoobscura* are not significantly substructured (SCHAEFFER and

MILLER 1992a). We present here an analysis of linkage disequilibrium for all segregating sites in the *Adh* region of *D. pseudoobscura* to determine how much association exists among polymorphic sites in a genomic sequence that has not been influenced by selection or population subdivision.

MATERIALS AND METHODS

Strains and nucleotide sequences: Ninety-nine sequences were used in this study of linkage disequilibrium. Ninety-one of the nucleotide sequences were determined previously (SCHAEFFER and MILLER 1991, 1992a,b). The population location, strain names, and the EMBL/GenBank Data Library accession numbers for these 91 nucleotide sequences are: Stemwinder Provincial Park, British Columbia, Canada: BN16, BC86, BC93, BC414, PS101, PS102, PS108, PS126, PS139, PS163, PS164, PS170 and PS192 (X62181–X62183, X62185, X62230-X62234, X64468-X64471); Rainbow Orchard, Apple Hill, California: AH43, AH162, AH135, AH133, AH122, AH69, AH100, AH144, AH54 and AH165 (M60979–M60988); Gundlach-Bundschu Winery, Sonoma, California: GB12, GB13, GB24, GB26, GB27, GB32, GB33, GB41, GB54, GB59, GB66, GB68, GB69, GB78, GB82, GB91, GB92, GB96, GB103, GB104, GB109, GB112, GB114, GB116, GB118 and GB119 (X62191–X62216); Bryce Canyon National Park, Utah: BR3F, BR5F, BR7F, BR8F and BR10F (X62186–X62190); Mesa Verde National Park, Colorado: MV21, MV27, MV28, MV36 and MV43 (X62220–X62224); Kaibab National Forest, Arizona: PS214, PS219, PS220, PS224, PS230, PS243, PS245, PS261, PS262, PS265, PS274, PS279, PS281, PS282, PS283, PS287, PS289 and PS297 (X64472–X64489); San Bernadino Mountains, California: SB2, SB6, SB14 and SB18 (X62235–X62238); Mexico: MT350, MT228, MXA31, MXA64 and MXZ26 (X62218, X62219, X62225-X62227); Brookings, Oregon: OREG3 and OREG4 (X62228, X62229); MacDonald Ranch, California: DPSE (Y00602); Baja California, Mexico: BJ855 (X62184); and Gregory Canyon, Colorado: GC4 (X62217).

This study also included eight new nucleotide sequences from strains collected from the Kaibab National Forest. This increased the sample size of the Kaibab population to 26 strains and the overall survey to 99 strains. SCHAEFFER and MILLER (1991, 1992a,b) describe the methods used to determine the nucleotide sequence of the *Adh* region of *D. pseudoobscura*. The names of the new strains and their GenBank/EMBL Data Library accession numbers are: PS298, PS299, PS302, PS306, PS309, PS314, PS315 and PS316 (X68159–X68166).

DNA sequence alignment: Figure 1 shows the fine structure of the *Adh* region as well as the fragment of DNA that was sequenced. The 99 nucleotide sequences were aligned manually with the Eyeball Sequence Editor [ESEE, version: 2.00a; (CABOT and BECKENBACH 1989)]. The alignments were determined by minimizing the number of mismatches and gaps assumed in the sequences. Any nucleotide site with two or more bases present in any population was defined as a segregating site or polymorphism. All segregating sites were used in linkage disequilibrium analyses. Insertion and deletion polymorphisms were excluded from all analyses and will be considered in later papers. A segregating site found within an insertion or deletion was also excluded from analyses of linkage disequilibrium.

Measures of linkage disequilibrium: We used five measures to estimate the degree of association between all pairs of segregating sites i and j in the *Adh* region D_{ij}, D^2, r_{ij}, r^2 and D'_{ij} [HUDSON (1985); see equations 1 through 5]. The number of independent estimates of linkage disequilibrium between two segregating sites is equal to $(n_i - 1)(n_j - 1)$ where n_i is the number of nucleotides segregating at site i and n_j is the number of bases segregating at site j (WEIR 1990). Fisher's exact test of independence was used to determine if the estimated linkage disequilibrium values were significantly different from zero (WELLS and KING 1980; SOKAL and ROHLF 1981). An association was considered significantly different from zero if the probability of the exact test was less than 0.05. This analysis had the potential to find significant nonrandom associations in excess of the significance level of the exact test. We adjusted our significance level with the sequential Bonferroni test procedure (RICE 1989), to overcome the multiple comparison problem.

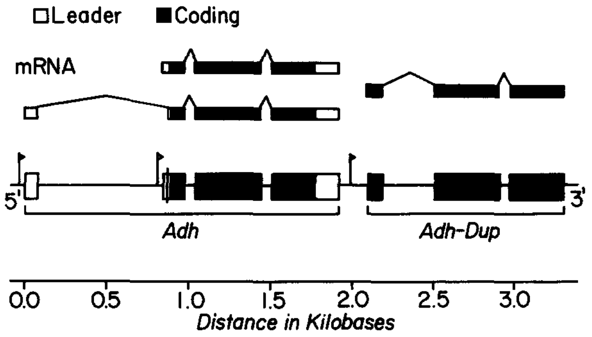

FIGURE 1.—Fine structure of the *Adh* region of *D. pseudoobscura*. The region is subdivided into 17 sequence domains: 5′ flanking, *Adh* adult leader, *Adh* adult intron, *Adh* larval leader, *Adh* exon 1, *Adh* intron 1, *Adh* exon 2, *Adh* intron 2, *Adh* exon 3, *Adh* 3′ leader, intergenic, *Adh-Dup* exon 1, *Adh-Dup* intron 1, *Adh-Dup* exon 2, *Adh-Dup* intron 2, *Adh-Dup* exon 3, and 3′ flanking.

Detection of clustered polymorphic sites: The STEPHENS (1985) test was used to detect sets of segregating sites that form congruent phylogenetic partitions. Sites that had the same phylogenetic information and were significantly clustered may indicate intragenic recombination or gene conversion. Lack of clustered polymorphisms suggests free recombination in the region. SAWYER (1989) also has a test to detect gene conversion, however, this method's test statistic may indicate that gene conversion has occurred but does not locate which sites are involved in the conversion event. Thus, Stephens test was used as a method to determine which segregating sites were involved in potential recombination or conversion events.

Estimates of nucleotide distance between segregating sites: The distance separating two segregating nucleotides may vary for different pairs of sequences due to the presence of insertions and deletions. The simplest method to estimate distance between a pair of sites is to subtract the 5′ nucleotide position number from the 3′ nucleotide position number which would include any insertions or deletions that occurred among all sequences. A more accurate method is to estimate the average number of nucleotides that separates the segregating sites across all sequences (Figure 2). These two measures of nucleotide distance are highly correlated ($R = 0.989$; $P < 0.01$). We used the former distance because it requires less time to estimate in computer simulations.

Relationship between linkage disequilibrium and nucleotide distance: Nucleotide sites that are tightly linked might be expected to be more associated than sites separated by larger nucleotide distances, *i.e.*, the distance effect. We determined if significant nonrandom associations decrease

```
  2<----------------- 28 Base Pairs ---------------->29
  ⇑                                                  ⇑ Actual
A G C T - G A A T - - - - - - T C C - - G C T A - T T T C G  16BP
A G C T - G A A T G A G C T T T C C - - G C T A - T T T A G  22BP
A G C T - G A A T - - - - - - T C C G C G C T A - T T T C G  18BP
A T C T T G A A T - - - - - - T C C - - G C T A - T T T C G  17BP
A G C T - G A A T - - - - - - T C C - - G C T A A T T T C G  17BP
                                                    Average  18BP
```

FIGURE 2.—Illustration of two methods to estimate distance between a pair of polymorphic sites. The top estimate of distance counts all positions between the pair of segregating sites including insertions and deletions. The distance estimate shown in the lower right is an average of the actual number of bases that separate the polymorphic sites in each sequence, shown along the right edge of the sequences.

as nucleotide distances increase by estimating correlation coefficients between nucleotide distance and each of the linkage disequilibrium estimators D_{ij}, D^2, r_{ij}, r^2 and D'_{ij} (SOKAL and ROHLF 1981). We used the absolute value of D_{ij}, r_{ij} and D'_{ij} in analyses of the relationship of linkage disequilibrium and distance. The statistical significance of each correlation coefficient was determined with two random permutation methods that shuffled the sites as columns in the aligned sequence. Let m be the total number of nucleotides in the aligned set of sequences and S be the number of segregating sites. Also, let m_i be the position of nucleotide site i in the aligned sequence (m_i, $i = 1, 2, 3, \ldots, m$) and S_i be the position of segregating site i in the aligned sequence (S_i, $i = 1, 2, 3, \ldots, S$). In this study, ($m =$ 3,618; $m_1 = 1$, $m_2 = 2$, $m_3 = 3, \ldots, m_{3618} =$ 3,618) and ($S =$ 359; $S_1 = 2$, $S_2 = 3$, $S_3 = 5, \ldots, S_{359} =$ 3,591).

The first method randomly shuffled the position of each nucleotide site, m_i, up to the total of m positions then the correlation between linkage disequilibrium and nucleotide distance was estimated for the permutation replicate. The second method randomly shuffled the positions of each segregating site, S_i, up to the maximum of S positions then the correlation between linkage disequilibrium and nucleotide distance was estimated for the permutation replicate. The fraction, P, of 1,000 permutation replicates with correlation coefficients more extreme than the observed value determined the statistical significance of the relationship. The first random permutation method assumes that all nucleotide sites are equally likely to mutate, while the second method assumes that not all nucleotide sites are free to mutate (*i.e.*, regulatory sequences or nonsynonymous sites in coding sequences).

Our linkage disequilibrium analysis showed two clusters of linkage disequilibrium (CLD) in the *Adh* region. These clusters add a large number of sites separated by short distances. To determine if these clustered sites generate a distance effect, we removed these sites from our analysis and repeated the correlation analysis described above for the five linkage disequilibrium estimators and two permutation methods.

Linkage disequilibrium and population structure: We examined the effect of population structure on levels of linkage disequilibrium by OHTA's (1982) method. OHTA defined variance components of linkage disequilibrium (D^2_{IS}, D^2_{ST}, D'^2_{IS}, D'^2_{ST} and D^2_{IT}; see equations 10 through 14, respectively) which are analogous to WRIGHT's (1940) inbreeding statistics. OHTA suggests that limited migration explains the observed linkage disequilibrium between populations if $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} > D^2_{IS}$. On the other hand, epistatic selection is more apt to explain the observed linkage disequilibrium between populations if $D'^2_{IS} < D'^2_{ST}$ and $D^2_{ST} < D^2_{IS}$. We examined the variance components of linkage disequilibrium for the Gundlach-Bundschu and Kaibab populations for the two clusters of linkage disequilibria in the introns of *Adh*. These two populations were used because the number of sequences collected from each location is 26. All other populations had sample sizes less than 10 individuals, thus, chi-square tests of homogeneity across all populations were not feasible.

Estimates of the mutation ($4N\mu$) and recombination ($4Nc$) parameters: We estimated the mutation parameter, $\hat{\pi} = 4N\mu$, by NEI's (1987; see equation 10.5) method, where N is the effective population size and μ is the neutral mutation rate. BROWN, FELDMAN and NEVO (1980) designed a set of summary statistics to describe levels of multilocus associations between genetic markers. They suggested that the moments of the distribution of K, the number of nucleotide differences between pairs of sequences, summarize levels of multilocus association. In particular, the variance of the random variable K can be represented as a function of pairwise linkage disequilibrium, D and D^2, summed over all pairs of sites. The variance of K may be used to estimate the recombination parameter, $\hat{C} = 4Nc$, from nucleotide sequence data if the variation in the region is selectively neutral (HUDSON 1987; see equations 1, 3 and 4), where N is the effective population size and c is the recombination rate between the ends of the region sequenced. $\hat{C}$ is inversely proportional to the average level of D and D^2 in the region.

We estimated $\hat{C}$, $\hat{\pi}$ and $\hat{C}/\hat{\pi}$ separately for *Adh* and *Adh-Dup* to determine if levels of recombination, mutation and the ratio of recombination and mutation differ between the two genes. *Adh-Dup* is a functional gene that resulted from an ancient duplication of the *Adh* gene or vice versa (SCHAEFFER and AQUADRO 1987; KREITMAN and HUDSON 1991). All segregating nucleotide sites were used to estimate $\hat{C}$ on a per locus basis. $\hat{C}$ and $\hat{\pi}$ were weighted by the number of base pairs in *Adh* or *Adh-Dup* to determine $\hat{C}$ and $\hat{\pi}$ on a per nucleotide basis. We used a bootstrap approach to determine a 95% confidence interval (CI) on the estimates of $\hat{C}$, $\hat{\pi}$, and $\hat{C}/\hat{\pi}$. For each replicate, a sample of 99 nucleotide sequences were drawn at random with replacement from the observed set of sequences. Sequences rather than sites were sampled at random because we wanted to simulate random draws of sequences from natural populations. We estimated $\hat{C}$, $\hat{\pi}$ and $\hat{C}/\hat{\pi}$ from the random draw of sequences in *Adh* and *Adh-Dup*. We randomly sampled the 99 sequences 1,000 times. The estimates of $\hat{C}$, $\hat{\pi}$ and $\hat{C}/\hat{\pi}$ for the 1,000 bootstrap replicates were each rank ordered and the 95% confidence interval was determined from the distributions of $\hat{C}$, $\hat{\pi}$ and $\hat{C}/\hat{\pi}$.

RESULTS

The distributions of linkage disequilibrium estimators: We observed 359 segregating sites in the 99 *Adh* sequences collected from the geographic range of *D. pseudoobscura*. Of the 359 polymorphic sites, 332 and 27 sites had two and three nucleotides segregating, respectively (Figure 3). We generated 74,278 independent values for each of the linkage disequilibrium estimators (54,946 estimates for all comparisons of 332 sites with two nucleotides; 1,404 estimates for all comparisons of 27 sites with three nucleotides; and 17,928 estimates for all comparisons of the 332 sites with two nucleotides and 27 sites with three nucleotides).

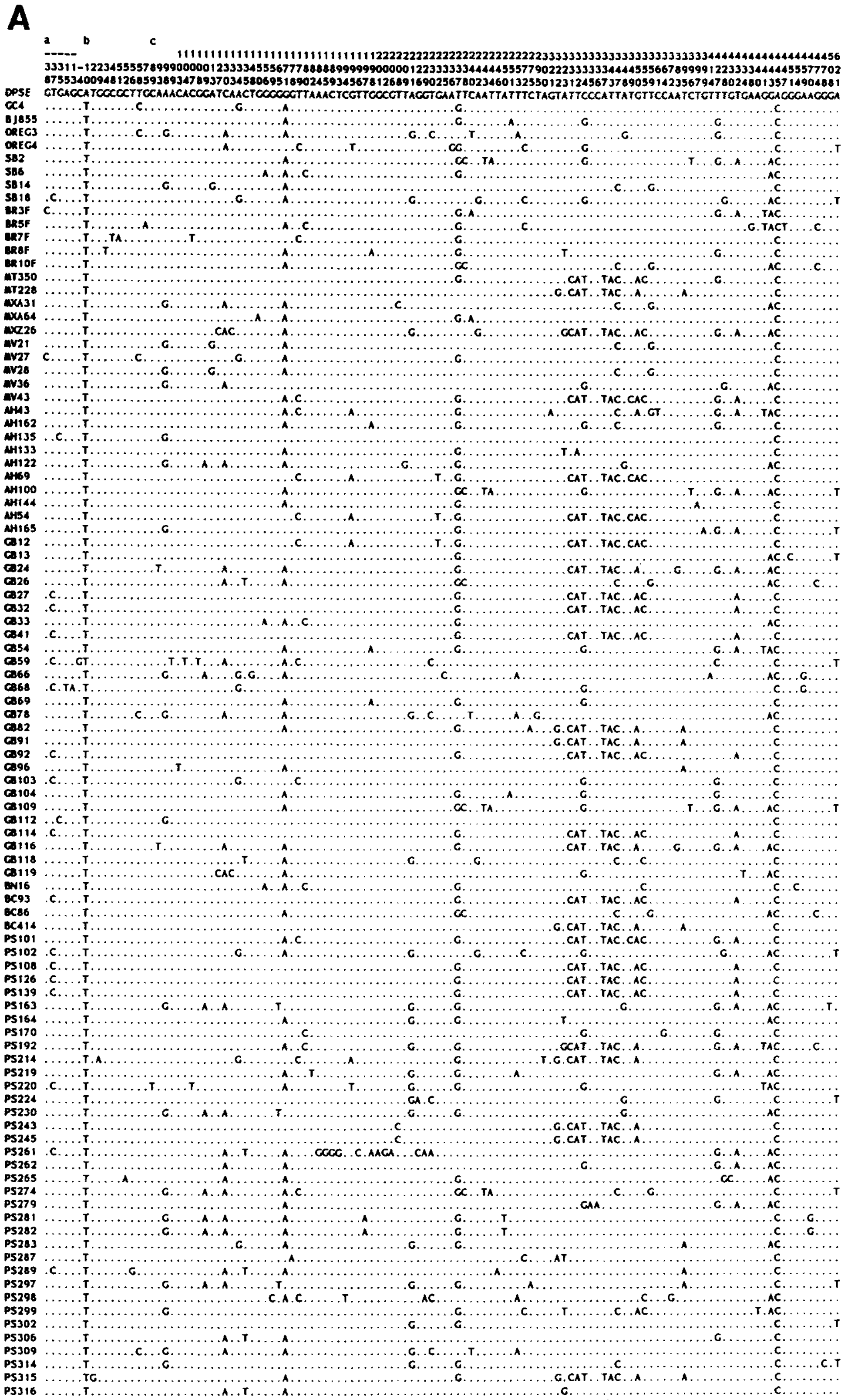

FIGURE 3.—Segregating sites found in 99 strains of *D. pseudoobscura* in the *Adh* region used in an analysis of linkage disequilibrium. The numbers above the sequence are the position numbers of each segregating sites within the reference sequence DPSE (SCHAEFFER and AQUADRO 1987). The sequence domains of *Adh* are indicated by letters and are shown above the site numbers. The domains are abbreviated as follows: a. 5′ Flanking ; b, *Adh* adult leader; c, *Adh* adult intron; d, *Adh* larval leader; e, *Adh* exon 1; f, *Adh* intron 1; g, *Adh* exon 2; h, *Adh* intron 2; i, *Adh* exon 3; j, *Adh* 3′ leader; k, intergenic; l, *Adh-Dup* exon 1; m, *Adh-Dup* intron 1; n, *Adh-Dup* exon 2; o, *Adh-Dup* intron 2; p, *Adh-Dup* exon 3; and q, 3′ flanking. The names of the sequences are given in MATERIALS AND METHODS.

The distributions for the five linkage disequilibrium estimators across all pairwise comparisons are shown in Figure 4. The theoretical distributions for D_{ij}, D^2, r_{ij}, r^2 and D'_{ij} have been determined by Monte Carlo simulation for two loci with known estimates of the mutation and recombination parameters, $4N\mu$ and $4Nc$ (GOLDING 1984; HUDSON 1985). The comparison of the observed and theoretical distributions should be viewed with caution because the distribution is based on comparisons of sites that differ in allele

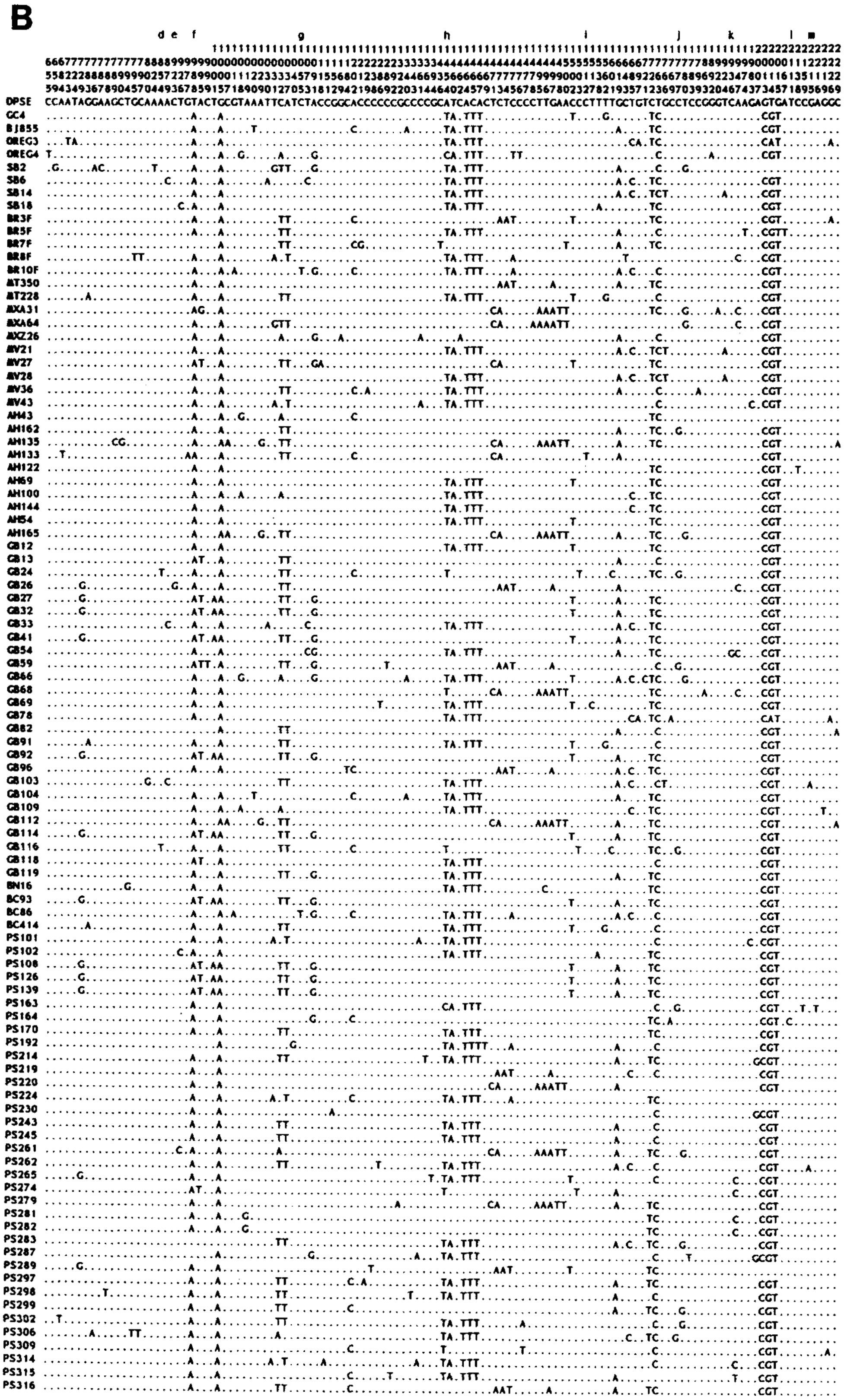

FIGURE 3.—Continued

frequency. In addition, the disequilibria values are correlated especially when one segregating site is in common. The observed distributions D_{ij} and r_{ij} were consistent with theoretical distributions that had used large values of $4Nc$ (HUDSON 1985) because much of the probability distribution is located at D_{ij} or r_{ij} equal zero. The observed distribution of D'_{ij} resembled the theoretical distributions that are based on $4Nc$ near zero because 90% of the D'_{ij} are either −1 or 1. The observed distributions of D^2 and r^2 had more values

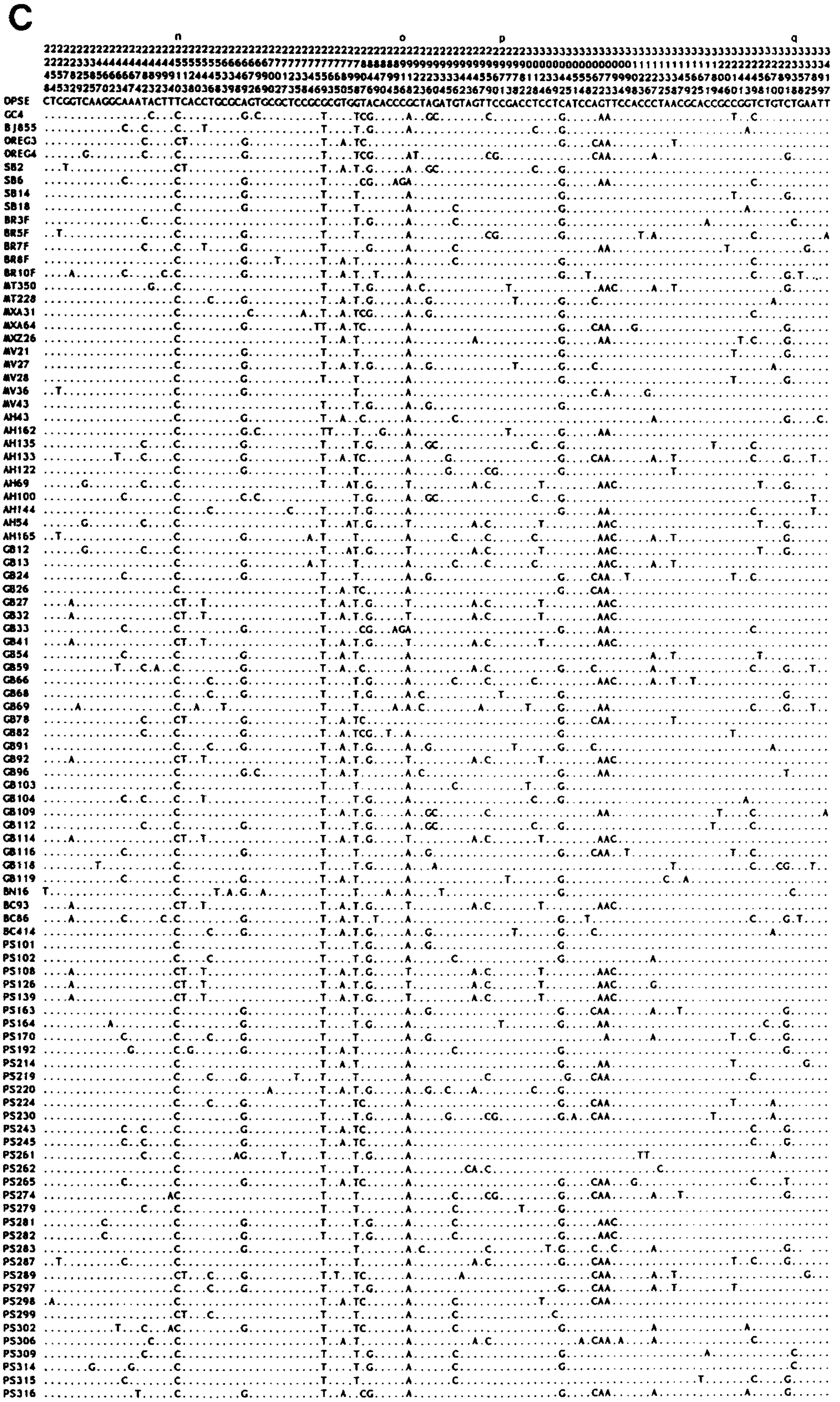

FIGURE 3.—Continued

near zero than the theoretical distributions with $4Nc$ equal to zero (GOLDING 1984; HUDSON 1985).

A total of 127 of the 74,278 (0.17%) pairwise comparisons of segregating sites were significantly different from zero when the Fisher's exact test was used with the sequential Bonferroni method (WELLS and KING 1980; SOKAL and ROHLF 1981; RICE 1989) (Figure 5). The overall significance level used was 6.7×10^{-7} (0.05 significance level/74,151 comparisons). Two clusters of segregating sites were significantly

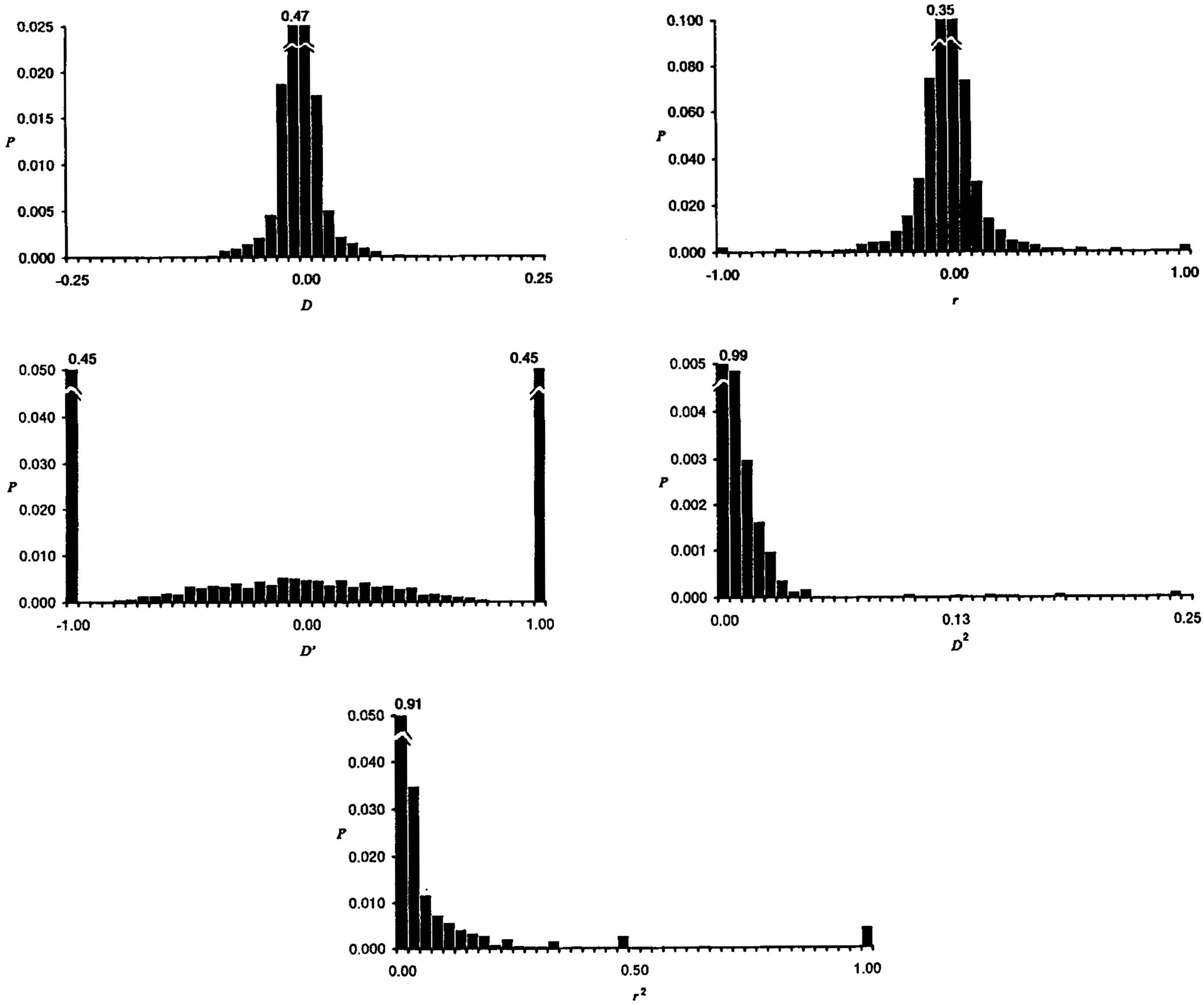

FIGURE 4.—A comparison of the distributions of five linkage disequilibrium estimators D_{ij}, D^2, r_{ij}, r^2 and D'_{ij}. The distributions are based on the comparison of all pairs of the 359 segregating sites in the *Adh* region.

associated in the *Adh* region, sites in the adult intron (nucleotides 331, 332, 334, 337, 343, 347, 350 and 355) and intron 2 of *Adh* (nucleotides 1454, 1460, 1464, 1465, 1467, 1471, 1473, 1474, 1475, 1495, 1496, 1497, 1498 and 1500) (Figure 5). These sites will be referred to as the clustered linkage disequilibrium (CLD) sites. Comparisons of sites between the two clusters were independently associated when tested with Fisher's exact method. Similar patterns of linkage disequilibrium were observed in the Gundlach-Bundschu ($n = 26$) and Kaibab ($n = 26$) populations (S. W. SCHAEFFER, data not shown).

The STEPHENS (1985) test detected 29 phylogenetic partitions with segregating sites that were significantly clustered. We observed 21 partitions that discriminated a single strain from the 98 others (Table 1) and eight partitions that grouped multiple strains together (Table 2), single and multiple strain partitions, respectively. The number of segregating sites within any of the phylogenetic partitions varied from a minimum of 2 to a maximum of 10 (Tables 1 and 2). The majority of phylogenetic partitions were composed of 2 or 3 polymorphic sites. The Stephens test detected some of the segregating sites in the CLD sites in the adult intron and intron 2 of *Adh* (Table 2; partitions 2 and 1, respectively). Some sites within the two CLDs were undetected by the Stephens test because not all the CLD sites have the same phylogenetic information.

Correlation of linkage disequilibrium and distance: The 127 pairs of sites in significant linkage disequilibrium were tested for the distance effect. The five linkage disequilibrium estimators were significantly negatively correlated with nucleotide distance when tested with both random permutation methods (Table 3). The Gundlach-Bundschu and Kaibab populations do not show a significant correlation between

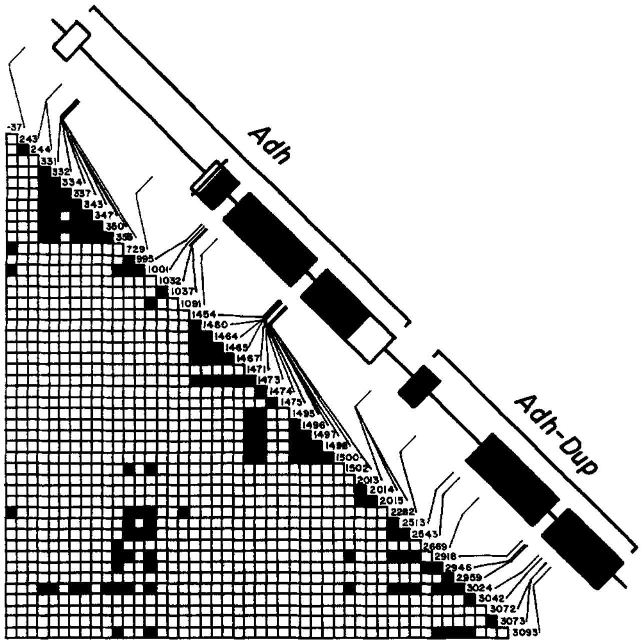

FIGURE 5.—Significant nonrandom associations among all pairs of 359 segregating sites in the *Adh* region in a sample of 99 strains of *D. pseudoobscura*. Each box in the matrix represents the comparison of two polymorphic sites. A black box represents a comparison where Fisher's exact test was significant at the 0.05 level. The location of the segregating sites relative to the two structural genes is shown on the diagonal. The only segregating sites included in the figure are those that show significant nonrandom associations.

nucleotide distance and any of the linkage disequilibrium estimators with either random permutation analysis (S. W. SCHAEFFER, unpublished data). We had little power to detect a distance effect in these populations because the number of strains and segregating sites was small. The distance effect disappears when the CLD sites were removed from the correlation analysis for four of the five estimates of linkage disequilibrium (Table 3). The distance effect still exists for D^2 when the CLD sites were removed.

Linkage disequilibrium and population structure: The two clusters of linkage disequilibrium may be due to either limited migration between demes or epistatic selection. We used OHTA's (1982) variance of linkage disequilibrium analysis to discriminate between the epistatic selection and limited migration hypotheses. All pairs of segregating sites within the adult intron and intron 2 were used to estimate D_{IS}^2, D_{ST}^2, $D_{IS}^{\prime 2}$, $D_{ST}^{\prime 2}$ and D_{IT}^2 from the Gundlach-Bundschu and Kaibab sequence data. The majority of comparisons showed D_{ST}^2 and $D_{IS}^{\prime 2}$ to be greater than D_{IS}^2 and $D_{ST}^{\prime 2}$, respectively, in the two intron sites (Figure 6). The data were more consistent with the epistatic selection hypothesis rather than the limited migration hypothesis. In other words, similar nonrandom associations were observed in the two populations. These data are consistent with the analysis of population structure by SCHAEFFER and MILLER (1992a).

TABLE 1

Clustered polymorphisms for single strain partitions in the *Adh* region

				Statistical significance	
Strain	S	d_o	g_o	$P(d \leq d_o)$	P
PS261	10	2653	2562	0.208	0.000
SB2	4	1725	1553	0.278	0.029
GB68	3	1943	1941	0.555	0.001
DPSE	3	2633	2179	0.818	0.001
PS315	3	3279	1958	0.975	0.001
PS214	3	1384	1095	0.327	0.001
BR8F	3	2801	1660	0.870	0.001
PS265	3	1391	994	0.330	0.001
MXA64	3	2726	1366	0.848	0.001
PS219	3	2984	2673	0.919	0.001
AH43	3	3195	3155	0.962	0.001
GB119	3	2840	2816	0.881	0.001
PS314	3	2020	1317	0.587	0.001
PS163	3	1789	1732	0.492	0.001
PS192	3	1572	1143	0.402	0.001
MXZ26	3	2157	1864	0.643	0.001
MXA31	3	867	801	0.145	0.002
PS164	3	1184	800	0.251	0.002
OREG3	2	1		0.001	
AH135	2	1		0.001	
PS306	2	40		0.022	

S, number of segregating sites; d_o, observed distance between the distal segregating sites; g_o, observed maximum distance between segregating sites for ($S \geq 3$); $P(d \leq d_o)$, probability of observing a distance less than d_o; and P, probability that a random segment of length, $d_o - (S - 1)$, has at least one segment greater than g_o.

Estimates of the mutation and recombination parameters in *Adh* and *Adh-Dup*: The estimate of $\hat{C}$ was significantly larger in *Adh-Dup* than in *Adh*. The estimate of $\hat{\pi}$ was significantly larger in *Adh* than in *Adh-Dup*. Our estimate of the ratio of the recombination and mutation rates, c/μ, was greater in *Adh-Dup* than in *Adh* (Table 4). The differences of $\hat{C}$, $\hat{\pi}$ and c/μ between *Adh* and *Adh-Dup* were statistically significant because the 95% confidence interval did not overlap for the two genes (Table 4).

HUDSON's (1987) estimate of $\hat{C}$ assumes that all the segregating sites are selectively neutral. OHTA's (1982) analysis suggests that the two clusters of segregating sites in the *Adh* introns may be in linkage disequilibrium because of epistatic selection. Thus, we estimated $\hat{C}$, $\hat{\pi}$ and c/μ after removing the CLD sites from the data set. The estimates of $\hat{C}$, $\hat{\pi}$ and c/μ were not statistically different between *Adh* and *Adh-Dup* when the CLD sites were removed (Table 4). The estimates of the ratio of recombination and mutation for the *Adh* region suggest that 7–17 recombination events have occurred for each mutation event (Table 4).

DISCUSSION

The distribution of linkage disequilibrium estimators: The expected value of linkage disequilibrium

TABLE 2

Clustered polymorphisms for multiple strain partitions in the *Adh* region

Partitions	S	d_o	g_o	Statistical significance $P(d \leq d_o)$	 P
1	4	7	4	0.000	0.187
2	3	12	6	0.000	0.174
3	2	6		0.003	
4	2	1		0.001	
5	2	2		0.001	
6	2	24		0.013	
7	2	3		0.002	
8	2	2		0.001	
All sites	359	3589	115	0.201	0.002

The symbols are defined as in Table 1. Strains that comprise each partition are as follows: Partition 1: DPSE, BR3F, BR7F, MT350, MXA31, MXA64, MXZ26, MV27, AH43, AH162, AH135, AH133, AH122, AH165, GB13, GB24, GB26, GB27, GB32, GB41, GB59, GB82, GB92, GB96, GB112, GB114, GB116, BC93, PS108, PS126, PS139, PS164, PS219, PS220, PS230, PS261, PS274, PS279, PS281, PS282, PS289, PS299, PS309, and PS316; Partition 2: MT350, MT228, MXZ26, MV43, AH69, AH54, GB12, GB24, GB27, GB32, GB33, GB82, GB91, GB92, GB114, GB116, BC93, BC414, PS101, PS108, PS126, PS139, PS192, PS214, PS243, PS245, and PS315; Partition 3: MXZ26 and GB119; Partition 4: SB2, AH100, GB109, and PS274; Partition 5: BR8F and PS306; Partition 6: AH135, AH165, and GB112; Partition 7: MXA31, MXA64; AH135, AH165, GB68, GB112, PS220, PS261, and PS279; and Partition 8: DPSE, SB2, AH43, PS224, and PS289.

TABLE 3

Correlation between nucleotide distance and five linkage disequilibrium estimators

Estimator	All sites included Correlation	 P_1	 P_2	All sites with CLD removed Correlation	 P_1	 P_2
D_{ij}	−0.32	<0.004	<0.005	−0.20	NS	NS
r_{ij}	−0.34	<0.004	<0.002	−0.05	NS	NS
D'_{ij}	−0.31	<0.002	<0.005	0.02	NS	NS
D^2	−0.34	<0.001	<0.001	−0.22	<0.041	<0.046
r^2	−0.38	<0.001	<0.002	0.00	NS	NS

P_1, probability of observing this correlation coefficient or a larger value determined from 1000 random permutations that shuffled the positions of all m nucleotides at random; P_2, probability of observing this correlation coefficient or a larger value determined from 1000 random permutations that shuffled the positions of the S segregating sites at random; and CLD, clusters of segregating sites in linkage disequilibrium in the adult intron and intron 2 of *Adh*.

for two loci in equilibrium populations is zero although its variance can be quite large (HILL and ROBERTSON 1968; OHTA and KIMURA 1969a). The comparison of all pairwise combinations of 359 segregating sites in the *Adh* region of *D. pseudoobscura* shows that 0.1% of the linkage disequilibria were significantly different from zero. The observed distributions of D_{ij}, D^2, r_{ij} and r^2 were each consistent with their theoretical distributions given our estimates of the recombination parameter (Figure 4, Table 4) (GOLDING 1984; HUDSON 1985). The observed distribution of D'_{ij}, however, differs significantly from its theoretical distribution with $\hat{C} = 20$ or ∞ presented in Figure 4 of HUDSON (1985). We found that 90% of the pairwise comparisons of sites estimated to be either −1.0 or 1.0. The frequency spectrum for the *Adh* region suggests an explanation for the observed U-shaped distribution of D'_{ij}. Of the 359 segregating sites in the region, 262 sites have frequencies of rare variants ≤0.05. These sites are likely to have estimates of D'_{ij} near −1.0 or 1.0 and to account for approximately 46% of the pairwise comparisons of the polymorphisms. Rare frequency sites are most likely to have arisen recently in the population by the introduction of new mutations. Therefore, the observed distribution of D'_{ij} may reflect linkage disequilibria due to recently introduced mutations.

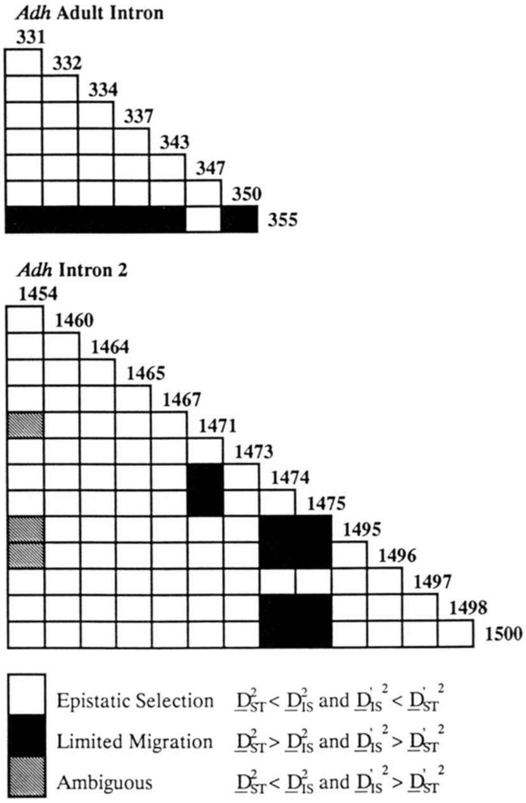
Adh Adult Intron
331
332
334
337
343
347
350
355
Adh Intron 2
1454
1460
1464
1465
1467
1471
1473
1474
1475
1495
1496
1497
1498
1500
Epistatic Selection $D^2_{ST} < D^2_{IS}$ and $D'^2_{IS} < D'^2_{ST}$
Limited Migration $D^2_{ST} > D^2_{IS}$ and $D'^2_{IS} > D'^2_{ST}$
Ambiguous $D^2_{ST} < D^2_{IS}$ and $D'^2_{IS} > D'^2_{ST}$

FIGURE 6.—Variance of linkage disequilibrium analysis for the Gundlach-Bundschu and Kaibab populations by OHTA's (1982) method. All pairwise comparisons of CLD sites within the adult intron and intron 2 of *Adh* were used to estimate D^2_{IS}, D^2_{ST}, D'^2_{IS}, D'^2_{ST} and D'^2_{IT}. Each box in the figure represents a comparison of a pair of sites. The position of each site in the standard sequence DPSE is shown along the diagonal. If a pairwise comparison finds $D'^2_{IS} < D'^2_{ST}$ and $D^2_{ST} < D^2_{IS}$ then the box is shown as an open box. If a pairwise comparison finds $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} < D^2_{IS}$ then the box is shown as a closed box. If a pairwise comparison finds $D^2_{ST} < D^2_{IS}$ and $D'^2_{IS} > D'^2_{IS}$ then the box is shown as a hatched box.

The distributions of the linkage disequilibrium estimators determined in this study should be viewed with caution because the distributions were deter-

TABLE 4

Estimates of the recombination and mutation parameters for the *Adh* and *Adh-Dup* genes

Gene	$4Nc$ (95% CI)	$4N\mu$ (95% CI)	c/μ (95% CI)
Adh	0.083 (0.047–0.097)	0.013 (0.012–0.013)	6.558 (3.749–7.742)
Adh-Dup	0.383 (0.135–0.453)	0.010 (0.010–0.011)	37.202 (13.323–44.583)
Total	0.087 (0.044–0.092)	0.012 (0.011–0.012)	7.453 (3.766–7.932)
Adh w/o CLD	0.184 (0.071–0.274)	0.009 (0.008–0.009)	21.283 (8.319–33.072)
Total w/o CLD	0.152 (0.061–0.155)	0.009 (0.009–0.010)	16.264 (6.644–16.925)

$4Nc$, neutral recombination parameter; $4N\mu$, neutral mutation parameter; CI, confidence interval; CLD, sites in clustered linkage disequilibrium in the adult intron and intron 2 of *Adh*.

mined from many segregating sites with different allele frequencies. In addition, some of the linkage disequilibria may be correlated when one segregating site is held in common. HUDSON's (1983a,b) simulation approach is a method that may determine the neutral expectation of linkage disequilibrium for an assemblage of segregating sites given our estimates of the mutation and recombination parameters. This approach would allow us to determine if clusters of linkage disequilibria develop with regularity and if an excess of rare frequency variants is typical under an infinite sites model of molecular evolution. Our estimates of $4N\mu = 37.83$ and $4Nc = 281.68$ (on a per locus basis) are prohibitively large for the computations in the HUDSON (1983a,b) simulation method to determine theoretical expectations.

The distance effect: We observed a significant distance effect. The distance effect disappears for four of the five linkage disequilibrium estimators when the two clusters of linkage disequilibrium in the *Adh* introns are removed from the analysis (Table 3). These data suggest that recombination is quite effective in breaking up associations among sites in this region. In fact, more recombination events occur relative to mutation events, 7 to 17. Thus, new linkage disequilibrium introduced by the mutation process is probably short lived in the *Adh* region of *D. pseudoobscura*.

The two clusters of linkage disequilibria in the introns of *Adh*: The only significant pattern of linkage disequilibrium observed in the *Adh* region were the two clusters of linkage disequilibria in the adult intron and intron 2 of *Adh*. Could the alignments generate these clusters of linkage disequilibrium? The adult intron cluster of 8 sites covers 25 base pairs of sequence; only 2 of the 8 segregating sites were consecutive and no gaps were assumed in the local alignment. The intron 2 cluster of 14 sites occupies 47 nucleotide positions; 2 and 4 segregating sites were in consecutive positions and one gap was assumed in the local alignment, however, the gap was not located near the consecutive bases. Commercially available sequence alignment programs (ALIGN DNASTAR, Madison, WI), with all their shortcomings, support the alignments that were used in these regions. In addition, not all the segregating sites in the clusters were in absolute or complete linkage disequilibrium, intermediate types were found. Therefore, we do not think that sequence alignments of the two regions are artificially creating linkage disequilibrium.

Three possible explanations could account for the clusters of linkage disequilibrium: (1) low recombination rates, (2) population subdivision or (3) strong epistatic selection. We have shown that estimates of the recombination rate in the *Adh* region are larger than the mutation rate because the ratio of the recombination and mutation parameters is significantly different from one (Table 4). Recombination rates do vary across the Drosophila genome (AGUADE, MIYASHITA and LANGLEY 1989a,b; STEPHAN 1989; STEPHAN and LANGLEY 1989; BEGUN and AQUADRO 1992), however, there is no evidence that recombination rates vary within a 4-kilobase region. Thus, the recombination rates that we estimate for the *Adh* region seem sufficient to break up the associations in the two intron clusters.

Population subdivision could create nonrandom associations among the intron sites. Two analyses suggest that *D. pseudoobscura* populations are not subdivided because gene flow is extensive between populations. SCHAEFFER and MILLER (1992a) have shown that at least two migrants are exchanged between populations each generation which is sufficient to homogenize gene frequencies between all North American populations of *D. pseudoobscura*. OHTA's (1982) variance of linkage disequilibrium analysis shows that most comparisons of sites in the two introns are more consistent with an epistatic selection hypothesis than with a limited migration or population subdivision hypothesis (Figure 6). In other words, the major haplotypes or multisite genotypes detected in the two intron clusters are found across all populations. Thus, population subdivision is an unlikely explanation for the clustered linkage disequilibrium.

If low recombination rate and population subdivision are unlikely explanations for the two intron clusters of linkage disequilibria then strong epistatic selection is left as the most probable hypothesis. LEWONTIN (1974) suggests that: "The observation of significant linkage disequilibrium that is consistent between populations is a very sensitive detector of natural selection." Why would the linkage disequilibrium analysis suggest epistatic selection and previous studies show no evidence of the past action of positive Darwinian selection (Schaeffer and MILLER 1992a,b)? The TAJIMA (1989) and HUDSON, KREITMAN and AGUADE (1987) tests of selective neutrality, performed in SCHAEFFER and MILLER (1992a,b), lumped nucleotide sites into two large regions *Adh* or *Adh-Dup*. Thus, tests of selective neutrality may have missed strong epistatic selection on two small local regions such as the intron clusters because the majority of sites in the region evolve under the assumptions of an infinite sites model of molecular evolution.

What is the nature of the epistatic selection? Analysis of pre-mRNA secondary structure of the *Adh* transcript indicates that the segregating sites in intron 2 form stem-loop structures (WOLFGANG STEPHAN, University of Maryland, personal communication). Ten of the segregating sites in intron 2 form three major haplotypes. Two of the haplotypes form stable stem-loop structures. The strong linkage disequilibrium among sites appears to be due to compensatory nucleotide changes that maintain mRNA secondary structure. We do not know whether the cluster of linkage disequilibrium in the adult intron of *Adh* is due to compensatory changes that maintain pre-mRNA secondary structure. Further theoretical and molecular analysis is necessary to confirm or refute the pre-mRNA hypothesis.

Analysis of linkage disequilibrium in the *Adh* region of *D. pseudoobscura* showed little evidence for nonrandom associations between segregating sites. Levels of recombination in the *Adh* region appear to be sufficient to rapidly breakup associations between nucleotide sites that are introduced by new mutations. The two clusters of linkage disequilibrium observed in the introns of *Adh* may result from strong epistatic selection to maintain secondary structure of the pre-mRNA.

We would like to thank CHARLES F. AQUADRO, ANDREW G. CLARK, CHARLES H. LANGLEY, KIMBERLYN NELSON, WOLFGANG STEPHAN, BRUCE S. WEIR, THOMAS S. WHITTAM and three anonymous reviewers for thoughtful comments on the analyses presented in this manuscript. This work was supported in part by grant GM42472 to S.W.S. from the National Institutes of Health.

Note added in proof: The two clusters of linkage disequilibrium observed in the *Adh* introns of *D. pseudoobscura* are similar to a pattern observed at the *white* locus of *D. melanogaster* (MIYASHITA and LANGLEY 1988). A recent analysis of the population structure of the cluster of linkage disequilibrium in the *white* locus suggests that strong epistatic selection maintains the nonrandom associations among segregating sites (MIYASHITA, AGUADÉ and LANGLEY 1993). These data may indicate a general phenomenon of nonrandom association that may be found at other genetic loci in the genome.

LITERATURE CITED

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989a Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989b Restriction-map variation at the *Zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 123–130.

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

BROWN, A. H. D., M. W. FELDMAN and E. NEVO, 1980 Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics **96:** 523–536.

CABOT, E. L., and A. T. BECKENBACH, 1989 Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. **5:** 233–234.

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

HILL, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. **8:** 117–126.

HILL, W. G., 1977 Correlation of gene frequencies between neutral linked genes in finite populations. Theor. Popul. Biol. **11:** 239–248.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

HUDSON, R. R., 1983a Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., 1983b Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

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

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KARLIN, S., and J. MACGREGOR, 1968 Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics **58:** 141–159.

KIMURA, M., and T. OHTA, 1971 *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton, N.J.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-Dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

MIYASHITA, N. M., M. AGUADÉ and C. H. LANGLEY, 1993 Linkage disequilibrium in the *white* locus region of *Drosphila melanogaster*. Genet. Res. (in press).

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

OHTA, T., 1982 Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:** 1940–1944.

OHTA, T., and M. KIMURA, 1969a Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63:** 229–238.

OHTA, T., and M. KIMURA, 1969b Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–55.

OHTA, T., and M. KIMURA, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

RICE, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

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

SCHAEFFER, S. W., and C. F. AQUADRO, 1987 Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics **117:** 61–73.

SCHAEFFER, S. W., and E. L. MILLER, 1991 Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogotá population of *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA **88:** 6097–6101.

SCHAEFFER, S. W., and E. L. MILLER, 1992a Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471–480.

SCHAEFFER, S. W., and E. L. MILLER, 1992b Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **132:** 163–178.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry*. W. H. Freeman and Co., New York.

STEPHAN, W., 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. II. The *Om(1D)* locus. Mol. Biol. Evol. **6:** 624–635.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. Genetics **121:** 89–99.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

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

SVED, J. A., 1968 The stability of linked systems of loci with a small population size. Genetics **59:** 543–563.

SVED, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. **2:** 125–141.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

WEIR, B. S., 1990 *Genetic Data Analysis*. Sinauer Associates, Inc., Sunderland, Mass.

WELLS, H., and J. L. KING, 1980 A general "exact test" for N × M contingency tables. Bull. South. Calif. Acad. Sci. **79:** 65–77.

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. Am. Nat. **74:** 232–248.

Communicating editor: B. S. WEIR