Inferring the Evolutionary Histories of the *Adh* and *Adh-dup* Loci in *Drosophila melanogaster* From Patterns of Polymorphism and Divergence

Martin Kreitman* and Richard R. Hudson†

**Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, and †Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717*

Manuscript received June 5, 1990
Accepted for publication November 27, 1990

ABSTRACT

The DNA sequences of 11 *Drosophila melanogaster* lines are compared across three contiguous regions, the *Adh* and *Adh-dup* loci and a noncoding 5′ flanking region of *Adh*. Ninety-eight of approximately 4750 sites are segregating in the sample, 36 in the 5′ flanking region, 38 in *Adh* and 24 in *Adh-dup*. Several methods are presented to test whether the patterns and levels of polymorphism are consistent with neutral molecular evolution. The analysis of within- and between-species polymorphism indicates that the region is evolving in a nonneutral and complex fashion. A graphical analysis of the data provides support for a hypothesized balanced polymorphism at or near position 1490, site of the amino acid replacement difference between Adh^f and Adh^s. The *Adh-dup* locus is less polymorphic than *Adh* and all 24 of its polymorphisms occur at low frequency—suggestive of a recent selective substitution in the *Adh-dup* region. Adh^s alleles form two distinct evolutionary lineages that differ one from another at a total of nineteen sites in the *Adh* and *Adh-dup* loci. The polymorphisms are in complete linkage disequilibrium. A recombination experiment failed to find evidence for recombination suppression between the two allelic classes. Two hypotheses are presented to account for the widespread distribution of the two divergent lineages in natural populations. Natural selection appears to have played an important role in governing the overall patterns of nucleotide variation across the two-gene region.

RECENT technical developments in molecular biology have led to a surge of interest in the use of gene frequency data to test theories of genetic variation. In *Drosophila melanogaster* alone, polymorphism levels have now been estimated using restriction enzymes (RFLP) for at least nine gene regions (*Adh*, see below; *87A Heat shock*, LEIGH BROWN 1983; *Amy*, LANGLEY *et al.* 1988; *Zw*, EANES *et al.* 1989; *rosy*, AQUADRO, LADO and NOON 1988; *yellow-achaete-scute*, AGUADÉ, MIYASHITA and LANGLEY 1989a, EANES *et al.* 1989, BEECH and LEIGH BROWN 1989; *zeste-tko*, AGUADÉ, MIYASHITA and LANGLEY 1989b; *white*, MIYASHITA and LANGLEY 1988; and *notch*, SCHAEFFER, AQUADRO and LANGLEY 1988); some of the same regions have also been studied in other *Drosophila* species (summarized in AQUADRO 1989). But the *Adh* region of *D. melanogaster* has received by far the most attention (LANGLEY, MONTGOMERY and QUATTLEBAUM 1983; KREITMAN 1983; AQUADRO *et al.* 1986; KREITMAN and AGUADÉ 1986a; AGUADÉ 1988; COLLET 1988; LAURIE-AHLBERG and STAM 1987, 1988; SIMMONS *et al.* 1989) and our understanding of the forces governing genetic variation within and between species comes almost exclusively from these studies (STEPHENS and NEI 1985; GOLDING, AQUADRO and LANGLEY 1986; KREITMAN and AGUADÉ 1986b; HUDSON, KREITMAN and AGUADÉ 1987; HUDSON and KAPLAN 1988).

Two RFLP studies, both based on four-cutter restriction enzymes, are sufficiently detailed to allow comparison of polymorphism levels between subregions of one contiguous stretch of DNA. One study identified 54 nucleotide polymorphisms across the *white* locus and its two flanking regions in *D. melanogaster* (MIYASHITA and LANGLEY 1988). A similar study of the *Xdh* locus (*rosy*) of *D. pseudoobscura* revealed 66 nucleotide polymorphisms in a 5.2 kilobase (kb) region (RILEY, HALLAS and LEWONTIN 1989). Neither study found evidence for heterogeneity of silent polymorphism levels among exons and noncoding regions. In contrast, a less detailed study of RFLP polymorphism in a 40 kb region containing the *rosy* and *snake* loci of *D. melanogaster* and *D. simulans* did show higher levels of polymorphism in the coding regions than in an upstream (and presumably noncoding) flanking region (AQUADRO, LADO and NOON 1988), but little is known about the extent of functional constraints in the noncoding region.

Of the nine gene regions studied in *D. melanogaster*, there is no convincing evidence of heterogeneity in polymorphism levels (KREITMAN 1990). One possible exception is the *yellow-achaete-scute* region of the *X* chromosome, which may be less polymorphic than the other eight regions. In a survey of 106 kb, only nine polymorphisms were detected at 176 restriction sites

in 64 chromosomes (AGUADÉ, MIYASHITA and LANGLEY 1989a). Reduced recombination at the distal tip of the *X* chromosome in *D. melanogaster* and the presence of selective substitutions is suggested as the cause of the low level of polymorphism, in accord with theory (KAPLAN, HUDSON and LANGLEY 1989). But, in the absence of a between-species comparison for evaluating the importance of selective constraint, this conclusion remains speculative. Also, three other studies of the same DNA region find substantially higher levels of variation (EANES *et al.* 1989; BEECH and LEIGH BROWN 1989; MACPHERSON, WEIR and LEIGH BROWN 1990). But these studies survey fewer sites and their higher polymorphism estimates are not statistically incompatible with the lower estimate.

In summary, restriction enzyme surveys have not allowed for critical evaluation of evolutionary theories pertaining to the maintenance of genetic variation. Most of these studies have surveyed too few sites to permit subregion comparisons, such as between introns, exons and noncoding regions. The inability to test for heterogeneity eliminates the possibility of applying even the most elementary tests for natural selection. Even comparing among loci, the very large variances of the heterozygosity estimates have limited the number of meaningful test comparisons. Furthermore, with the exception of *Adh,* none of the studies yet include between-species comparisons. Such a comparison is necessary to evaluate the contribution of selective constraint to variation in levels of polymorphism.

Therefore, *Adh* remains the only locus in Drosophila for which selective neutrality has been rejected as an explanation for the patterns of polymorphism within species and divergence between species. Specifically, HUDSON, KREITMAN and AGUADÉ (1987) rejected a neutral model using a conservative test of the predicted relationship between levels of polymorphism and divergence under neutrality (hereafter called the HKA test). By including between-species comparisons, the test takes into account between-region differences in the proportion of sites subject to selective constraint. Polymorphism estimates for that study were based on a four-cutter RFLP survey of 60 chromosomes derived from a single natural population and divergence data were based on a DNA sequence comparison of one *D. melanogaster* and one *D. sechellia* allele.

Although the test does not indicate the cause of departure from selective neutrality, the coding region of *Adh* contains a greater number of silent polymorphisms than the neutral prediction. Such a pattern is expected for neutral variation linked to a balanced polymorphism (STROBECK 1983). A reanalysis of the *Adh* data by HUDSON and KAPLAN (1988), who modelled the evolution of neutral polymorphisms linked to a balanced polymorphism, showed the pattern of silent variation to be compatible with a model assuming a balanced polymorphism at the site encoding the Adh^f and Adh^s alleles.

In this paper we expand the analysis of eleven *Adh* alleles by presenting the DNA sequences of two more regions, a 1350 base pair (bp) region located directly upstream of *Adh* and a 1300 bp downstream region. The latter region contains the complete presumptive coding sequence of *Adh-dup*, a functional gene that is distantly related to *Adh* by tandem duplication (SCHAEFFER and AQUADRO 1987). The data allow us to use sequence rather than RFLP comparisons for the HKA test and to test the *Adh* coding region for deviations from selective neutrality against both 3′ and 5′ regions. Assuming a balanced polymorphism for Adh^f and Adh^s, we would predict HKA test departures from neutrality for the *Adh* locus *vs.* either the 5′ flanking region or the 3′ *Adh-dup* locus but not necessarily for the 5′ flanking region *vs. Adh-dup*. We also present refinements to methods for exploring polymorphism and divergence data and for testing evolutionary hypotheses. Attention is paid to the problem of inferring the evolutionary forces acting in a region when there is more than one force.

MATERIALS AND METHODS

Fly stocks: Eleven isogenic lines for the second chromosome, consisting of five Adh^f and six Adh^s alleles from five population samples, are described in KREITMAN (1983). The *D. simulans* sequence is from a complete 4.6-kb sequence (COHN and MOORE, 1988; GenBank In: Drsadha). Sequences were manually aligned to minimize the total number of substitutions. There is essentially no ambiguity about the alignment of the sequences because the species, being closely related, differ at only a small percentage of sites (COYNE and KREITMAN 1986).

Sequence analysis: Cesium chloride gradient-purified genomic DNA was used as substrate for asymmetric polymerase chain reaction (PCR) amplification (KREITMAN and LANDWEBER 1989). PCR amplification and DNA sequencing used 20 base oligonucleotide primers that were purified either by gel electrophoresis or thin-layer chromatography. Primers for dideoxy sequencing were spaced at approximately 300 bp intervals. Dideoxy sequencing was performed using [^{35}S]dATP (Amersham) and Sequenase modified T7 polymerase (U.S. Biochemicals) according to directions supplied by the manufacturer. Complete sequences of both DNA strands were determined for each allele. The two strands always produced complimentary sequences with a single exception of one base difference in one line. However, the conflicting base sequence could not be reproduced. This was the only artifact that could possibly have been associated with the PCR method.

RESULTS

Distribution of polymorphism: The downstream gene, which we call *Adh-dup* for convenience, is related to *Adh* by an ancient tandem duplication. *Adh-dup* consists of three nonoverlapping open reading

TABLE 1

***Adh* 5′ flanking polymorphism**

Site:	1		11		21		31	
Reference:	G A.AAC	A ACG.G	T AA.AC	.CACT	CAGCA	G A ATAΔC	TTTCC	G
Wa^s	..G..	..T.A	.T...	Δ.C..	..T.G	...1.	.....	.
Fl^{1s}	T...A	Δ.A..	..ΔG.	...T.	..T.G	...1.	GAA..	C
Af^s	T...A	Δ.A..	..ΔG.	...T.	....G	G..2A	.....	.
Fr^s	..GTA	.GT.A	.T...	..C.A	TG.T.	.AT3.	GAA..	.
Fl^{2s}	T...A	Δ.A..	..ΔG.	...T.	....G	...1.	.....	.
Ja^s	..G..	..T.A	.T...	..C..	..T.G	...1.	GAA..	C
Fl^f	..G..	..T.A	.T...	..C..	..T.G	...1.	GAA..	C
Fr^f	.....	..T.A	TT.GT.	ΔTC..	....G	G..3.	...TG	C
Wa^f	.....	..TΔA	.T.GT.	ΔTC..	....G	G..3.	...TG	C
Af^f	.Δ...	..T.A	.T.GT.	ΔTC..	....G	G..4.	...TG	C
Ja^f	T...A	Δ.A..	...G.	...T.	....G	...1.	...TG	C

The reference sequence is the consensus of *D. simulans*, *D. mauritiana* and *D. sechellia* sequences (COYNE and KREITMAN 1986). If the outgroup sequences were polymorphic both sequences are given. Δ = insertion/deletion. Nucleotide positions corresponding to site Nos. are given in the APPENDIX (Table 8) 1, TOA; 2, TTT; 3, OAA; 4, TAA (where O = deleted).

frames that are conserved between *D. pseudoobscura* and *D. simulans* (SCHAEFFER and AQUADRO 1987). The same open reading frames occur in the *D. melanogaster Adh-dup* sequence presented here, except that the deduced polypeptide has one additional carboxy-terminal amino acid. A polyadenylated messenger RNA from adult flies can be identified by Northern analysis (FAGLES 1989) and the positions of two introns have been confirmed by PCR analysis of total RNA (DENKER 1990). From these observations there can be little doubt that *Adh-dup* encodes a functional protein and that our assignment of the coding region is correct.

A complete DNA sequence of the *Adh* 5′ flanking region, the *Adh* locus and *Adh-dup* is presented in Figure 9 (APPENDIX). The numbering system is the same as that used by KREITMAN (1983), which begins numbering at the initiation of transcription of the "adult" mRNA. The observed sequence differences among the 11 lines are given in Tables 1, 2 and 3 for the 5′ flanking, *Adh* and *Adh-dup* regions, respectively and their site locations are given in Table 8 (APPENDIX).

The 11 sequences contain 98 polymorphic sites, 36 in the 5′ flanking region, 38 in *Adh* and 24 in *Adh-dup*. Fourteen of the 98 polymorphisms are insertions or deletions; they are excluded from the following analysis. The distribution of silent and replacement polymorphism is presented in Figure 1, which also includes the distribution of silent and replacement differences between one *D. melanogaster* allele (Af^s) and a *D. simulans* allele (In: Drsadha). The positions of the between-species differences are shown in Figure 9 and are listed by site in Figure 10. Most of the differences between species are fixed differences, so the choice of a particular allele has little influence on the pattern or level of divergence. The sample contains one amino acid replacement polymorphism in *Adh-dup*, a charge-preserving isovaline to leucine difference. One amino acid polymorphism is also found in *Adh*, the lysine to threonine change at position 1490 that distinguishes Adh^s and Adh^f.

Differences between alleles: Two alleles, $Fl1^s$ and Wa^s (hereafter referred to as the Wa^s lineage or allele), are distinguishable from the remaining four Adh^s alleles at a total of nine silent sites in Exon 3, Intron 3 and Exon 4 of *Adh* and at a total of seven silent sites in Intron 2 of *Adh-dup*. The two lineages also differ by two insertion/deletions in *Adh-dup* and a single difference in the 3′ noncoding region of *Adh-dup*. Eighteen of the 19 mutations are in complete linkage disequilibrium. Using the *D. simulans* sequence as an outgroup for the purpose of assigning polarity, eleven mutations can be shown to have occurred on the Wa^s lineage and seven on the "standard" Adh^s lineage, as shown in Figure 2. Interestingly, none of the 36 polymorphic sites in the *Adh* 5′ flanking region are completely correlated with the Wa^s lineage; the distinction between the Wa^s and standard Adh^s lineages in this region may have been obliterated by recombination.

The Wa^s allele is distinguishable from the standard Adh^s allele at five sites by four-cutter RFLP analysis (SIMMONS *et al.* 1989). This RFLP haplotype is consistently found at low frequency in U.S. east coast populations (SIMMONS *et al.* 1989; A. BERRY and M. KREITMAN, unpublished data) and, with few exceptions, all five distinguishing site differences are in complete linkage disequilibrium. Therefore, evolutionary intermediates between Wa^s and Adh^s alleles are either rare or absent.

The large number of mutational differences and the lack of recombinants (or even evolutionary intermediates) between the two alleles suggest they have

TABLE 2

Adh Polymorphism

Site:	1 In1	14 Ex2	In2	18 Ex3	22 In3	27 Ex4 * C	36 3′
Reference:	CGATAAGGG.C.G	CT	AC	CTTCC	CGATT	CTCCACCAG	C.C
Wa^s	.A....T.....	..	..	T..A	.A..A	AC....T..	A..
Fl^{1s}	.A...T.....	..	..	T..A	.A..A	AC....T..	A..
Af^s	.A...T.....	..	..	.C..	G..A.	......T.A	A..
Fr^s	.A...T.....	..	GT	.C..	G..A.	......T.A	A1.
Fl^{2s}	A......TC....	AG	GT	.C..	G..A.	......T..	.3.
Ja^s	.A...T.....	.G	..	.C..	G..A.	...T.TTCA	.4.
Fl^f	.A...T.....	.G	..	.C..	G..A.	..GTCT.C.	.4.
Fr^f	A......TCΔGΔ.	.G	..	.C..	G..A.	..GTCT.C.	.4G
Wa^f	A......TCΔGΔ.	.G	..	.C..	G..A.	..GTCT.C.	.4G
Af^f	A......TCΔGΔ.	.G	..	.C..	G..A.	..GTCT.C.	.5G
Ja^f	A.GGG....Δ..T	.G	..	.CA.	G..A.	..GTCT.C.	.4.

* = LYS ↔ THR.

TABLE 3

Adh-dup polymorphism

Site:	1 5′	4 In1	9 Ex2	11 In2	19 Ex3	23 3′
Reference:	TGC	.C.T.	GA	CACT.TCT	AGGG	GA
Wa^s	A..	....Δ	..	..T.ΔA.G	....	A.
Fl^{1s}	A..	....Δ	..	..T.ΔA.G	....	A.
Af^s	A.T	Δ.1A.	AT	AG.A..G.	...A	..
Fr^s	.A.	.....	..	AG.A..G.	....	..
Fl^{2s}	A..	.....	..	AG.A..G.	....	..
Ja^s	A..	.T...	..	AG.A..G.	....	.Δ
Fl^f	A..	.....	..	AG.A..G.	GAT.	A.
Fr^f	A..	.....	..	AG.A..G.	....	..
Wa^f	A..	.....	..	AG.A..G.	....	..
Af^f	A..	.....	..	AG.A..G.	....	..
Ja^f	A..	.....	..	AG.A..G.	....	..

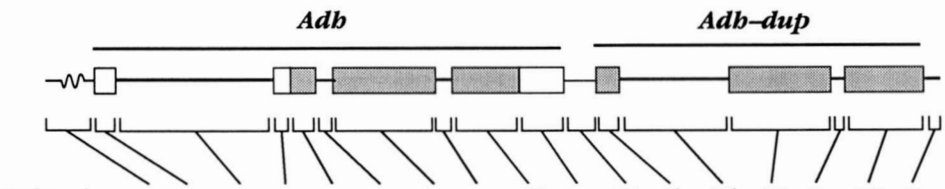
* = VAL ↔ ILE.

evolved in isolation of one another. Two models of isolation can be envisioned: geographic and genetic. Under the geographic model *D. melanogaster* is presumed to have subdivided into two isolated (or semi-isolated) populations which diverged one from another at the Wa^s sites. Since the two alleles are currently found together in natural populations, mixing of the subpopulations subsequent to divergence must be hypothesized. The mixing might have occurred during the expansion of the species from Africa (DAVID and CAPY 1988). This hypothesis makes two predictions. First, the subdivided populations may still exist somewhere in the species range, perhaps in the western African home-range (LEMEUNIER *et al.* 1986). Second, if the subdivision is geographic then the whole genome should have diverged. Therefore, similarly diverged lineages should be found at other places in the genome.

Under the genetic isolation hypothesis the two alleles evolved in isolation by the presence of recombination suppression. Inversions, for example, can suppress recombination in inversion heterozygotes.

We have tested for recombination suppression with two lines representing each allelic class, Wa^s and Af^s, by asking whether either allele suppresses recombination between two recessive visible markers located close to but on either side of *Adh*. Crosses were made between a marker stock, *b el* (*Adh*) $l(2)br3^{L692}$ rd^s *pr cn*/In(2LR)O, *Cy* dp^{wl} *pr cn* (WOODRUFF and ASHBURNER 1979) and either Wa^s or Af^s and F_1 females were backcrossed with the marker stock. $l(2)br3^{L692}$ is the closest recessive lethal complementation group proximal to *Adh*. *Elbow* is located at position 50.0 on the genetic map, *Adh* at 50.1 and $l(2)br3^{L692}$ also at 50.1 but definitely proximal to *Adh*. We determined the frequency of black and elbow F_2 flies, *e.g.*, ones indicating a single recombination in the region spanned by *el* and $l(2)br3^{L692}$. All crosses were carried out at 25°. A total of 10 black elbow and 7294 wild-type flies were recovered from the Af^s crosses and 5 black elbow and 7658 wild-type flies were recovered from the Wa^s crosses. These values are consistent with a predicted map distance between *el* and $l(2)br3^{L692}$ of 0.1 cM and the difference between the two strains is not statistically significant (χ^2 with continuity correction = 1.269, probability = 0.26). Therefore, we find no evidence of strong recombination suppression in either class of alleles.

Tests of neutrality based on distribution of polymorphism: First, we would like to know whether the distribution of polymorphism across regions is heterogeneous. We anticipate four "causes" of heterogeneity: (1) variation in the mutation rate, (2) variation in selective constraint, (3) genetic drift, and (4) natural selection at one nucleotide site influencing polymorphism levels at surrounding linked sites. We are not specifically interested, in this study, whether regions of different functionality exhibit different levels of polymorphism. Indeed, we would like to control for such differences and focus only on any remaining



Total Length:	1243	87	620	70	99	65	405	70	264	178	124	96	376	405	51	315	58
Effective #																	
Silent Sites:	1243	87	616	70	25.5	61	103.6	66	63.1	175	124	20.7	372	91.9	47	68.2	55
Within mel.:																	
Silent:	30	0	11	1	1	2	4	5	8	2	3	0	2	2	7	3	1
Repl.:	--	--	--	--	0	--	0	--	1	--	--	0	--	0	--	1	--
Mel. vs sim.:																	
Silent:	77	1	19	2	2	1	3	5	5	6	3	1	18	13	8	10	2
Repl.:	--	--	--	--	1	--	1	--	0	--	--	1	--	4	--	1	--

FIGURE 1.—Distribution of nucleotide variation in eleven lines of *D. melanogaster* and between *D. melanogaster* (Af^s) and *D. simulans* (In: Drsadha). The structure proposed for *Adh-dup* is based on conservation of open reading frames (as described).

		Adh	*Adh-dup*
Reference	Adh^s	CCAGGATAC	ΔAGCAΔTGTG
	Wa^s	TTACATAAC	ΔCATTΔACGA
	Af^s	.....	.....
Reference	*D. sim.*	CTCCGTTCT	ΔCACTΔTGTG
	Wa^s	T.A.A.AAC	Δ..T.ΔA.GA
	Af^s	.T.C.T...	.AG.A..G..

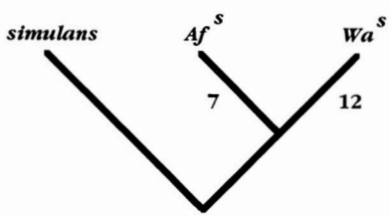


FIGURE 2.—Cladistic analysis of the nineteen sites distinguishing the Wa^s and standard Adh^s. *D. simulans* is the outgroup.

heterogeneity. It is this residual heterogeneity that may contain evidence for natural selection.

What we would like to know, then, is whether polymorphism levels still vary after mutation, selective constraint and genetic drift are all taken into account. This means that we will not contrast polymorphism levels in regions of different functionality, such as between introns and exons. And because natural selection is expected to affect polymorphism levels within physical blocks rather than among functional domains, we should avoid grouping together spatially separated regions of similar functionality for the purpose of comparing levels of polymorphism. To accomplish this goal we will restrict our attention to the distribution of only silent and noncoding polymorphism and will ignore, for now, all amino acid replacement sites and certain other sites known to be absolutely conserved (the intron splice junction sequences, "GT/CA" and "AC/TG").

The possibility of directional change in mutation rates affecting certain regions and not others may not be separable from other factors and cannot be specifically addressed from polymorphism data alone. So for the purpose of the present analysis, we assume the rate to be a constant. (This assumption will be relaxed in the next section, which considers interspecific divergence.)

A certain amount of heterogeneity is expected to arise from selective constraint. To control for this effect, our strategy is to compare only silent polymorphism levels. Our naive hypothesis is that, on average, the intensity of selective constraint at silent sites is the same across all DNA regions. All sites that, *a priori*, are expected to be selectively constrained are removed from consideration. First, to control for obvious constraints in coding regions we consider only "silent" positions (see KREITMAN 1983). Silent sites are defined as positions which when mutated do not change a polypeptide sequence. The number of silent sites in a region is the number of possible silent nucleotide changes divided by three (the number of possible changes at each site). The first and last two bases in introns, "GT" and "AC," necessary for splicing, are also considered nonsilent. All other noncoding regions in the subsequent analysis are assumed to contain only silent sites.

A certain number of noncoding sites are expected to contain critical and conserved regulatory sequences, but these sites will have a small effect on average polymorphism levels so long as they represent a sufficiently small proportion of the noncoding sites. It is important to point out that "silent" does not necessarily imply selectively neutral; a low level of constraint is possible at silent sites.

Before testing for heterogeneity across regions a decision must be made about how to subdivide the 4500-bp stretch of DNA. Particular attention must be paid to the "scale" of heterogeneity. For example, heterogeneity on the scale of 250 bp might not be detected for data subdivided into 1000-bp regions.

We are particularly interested in detecting silent polymorphism heterogeneity that is induced by natu-

ral selection acting at a linked site. The interval size that is appropriate for detecting selection must depend on the scale of selective effects. Unfortunately, different kinds of selection have different scales of effects. For example, a balanced polymorphism is expected to elevate neutral polymorphism levels but the size of the region depends on relative magnitudes of the mutation and recombination rates, the population size and the population frequency of the site under selection. For *D. melanogaster*, this region might be only 100 bp (HUDSON and KAPLAN 1988). Other kinds of selection, such as adaptive substitution, can produce a large range of lengths over which the neutral polymorphism level can be reduced. With directional selection the size of the region affected depends on the relationship of many factors, including the recombination and mutation rates, the strength of selection and the time since the selection event.

In the absence of any strong prediction about the scale of heterogeneity to be expected under selection, we have chosen to investigate three scales. The scales represent a balance between statistical considerations (there are only 81 silent polymorphisms) which impose practical limits on the number of subdivisions, and the recognition that the DNA has ordered functional subdivisions. The first and coarsest scale compares the *Adh* 5′ flanking region, the *Adh* locus and the *Adh-dup* locus. A finer scale, with five regions, subdivides the *Adh* and *Adh-dup* loci into an *Adh* 5′ noncoding region, an *Adh* coding region, a noncoding region containing the *Adh* 3′ nontranslated and the *Adh-dup* 5′ flanking sequences, and the *Adh-dup* coding region. As an alternative to this scheme, we also compare five regions, each containing the same number of silent sites, but which do not respect any functional boundaries.

To ask whether polymorphism levels are statistically different we must specify a model incorporating genetic drift. One simple model, which we now consider, is the infinite sites neutral model. Mutations are assumed to occur at random along the sequence. The number of sites is assumed to be large and the number of mutations small, so that sites are mutated no more than once in the population. The distribution of mutations per replication is, therefore, Poisson. Mutations are also assumed to have no effect on fitness; they are selectively neutral. A new generation is produced from the previous generation according to the Wright-Fisher model [see EWENS (1979) for details]. This is a discrete generation model in which, under diploidy, $2N$ gametes—possibly mutated—are sampled from N individuals to produce the next generation of zygotes. The distribution of offspring number is also, approximately, Poisson.

Under the infinite sites model polymorphism is governed by a single parameter, $\theta_T = 4N\mu_T$, where N is the population size and μ_T is the total mutation rate per generation at the locus. To apply this model to our data, any collection of silent sites can be approximated as an infinite site neutral process in which $\theta_T = m\theta$, where m is the number of silent sites and θ is a constant of proportionality which can be interpreted as $4N\mu$, where μ is now the mutation rate per silent site per generation. Our null hypothesis under this model is that the parameter θ is the same for all sites and all regions. An unbiased estimator of θ is given by

$$\hat{\theta} = s_n a_{n-1}^{-1} m^{-1} \quad \text{where} \quad a_{n-1} \sum_{i=1}^{n-1} 1/i \qquad (1)$$

where m is the number of sites under consideration and s_n is the number of segregating sites in the sample of n genes. The variance of θ is bounded by

$$\theta\, a_{n-1}^{-1}$$

for free recombination, and

$$\theta\, a_{n-1}^{-1} + \theta^2\{1 + 1/4 + \ldots + (n-1)^{-2}\}a_{n-1}^{-2} \qquad (2)$$

for no recombination (WATTERSON 1975).

Estimates of θ are given in Table 4 for the three regions and also for the introns and coding regions. The variances for these estimates are large, as shown in the table. Since the estimator θ is not normally distributed it is not appropriate to assign a confidence interval by adding and subtracting two standard deviations. To get a better idea of the range of values of θ that are compatible with each estimate we calculated the smallest (θ_L) and largest (θ_U) values of θ that are compatible at the 0.05 probability level with the observed numbers of polymorphic sites in each region. This was done by setting the following recursion equation for the probability, $P_n(s)$, of observing s segregating sites in a sample of size n to the desired probability (HUDSON 1990),

$$P_n(s) = \sum_{i=0}^{s} P_{n-1}(s-i)Q_n(i) \qquad (3)$$

where

$$Q_n(s) = \left(\frac{m\theta}{m\theta + n - 1}\right)^s \frac{n-1}{m\theta + n - 1}.$$

To obtain θ_U the following equation was solved for θ by computer iteration:

$$0.025 = \sum_{i=0}^{s_{obs}} P_n(i)$$

The values of θ_L were obtained by solving the following for θ:

$$0.975 = \sum_{i=0}^{s_{obs}-1} P_n(i).$$

The range of θ values compatible with observed polymorphism levels, shown in Figure 3, give a clear indication of the large stochastic and sampling vari-

TABLE 4

Estimates of the neutral parameter θ for three regions

Region	No. sites	No. polymorphic	$\hat{\theta}$	Var $(\hat{\theta})$ $(\times 10^5)$	θ_L	θ_U
5′ Flanking	1243	30	0.008	0.2–1.5	0.004	0.019
Adh	1267.2	34	0.009	0.2–1.8	0.005	0.021
Intron 1	616	11	0.006	0.3–1.0	0.003	0.016
Introns 2 + 3	127	7	0.018	5.0–11.5	0.007	0.054
Coding	192.2	13	0.023	4.1–13.8	0.010	0.058
Adh-dup	655.8	18	0.009	0.5–2.1	0.005	0.023
Introns 1 + 2	419	9	0.007	0.6–1.6	0.003	0.020
Coding	180.8	5	0.009	1.8–3.4	0.003	0.029

θ_L and θ_U are lower and upper critical values of θ for which there is a 5% probability of observing the same number or more extreme values for number of polymorphic sites.

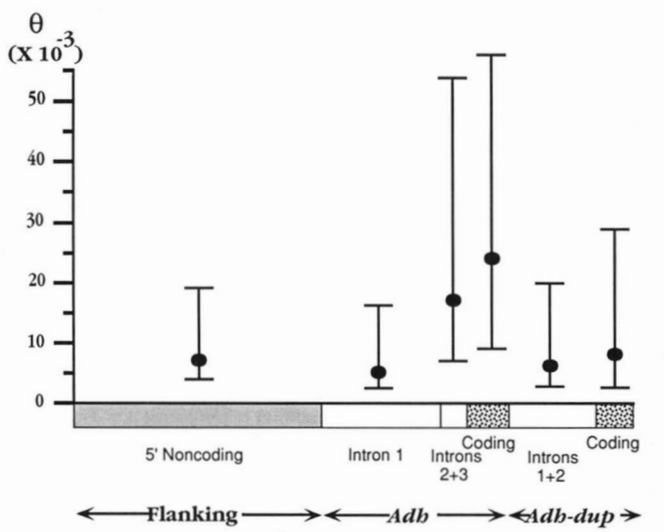


FIGURE 3.—Estimates of θ for several regions. 95% confidence intervals were calculated using Equation 3. The upper and lower limits define values of θ for which there is a 2.5% probability of observing the same number or greater (or the same number or fewer) polymorphisms than the observed number.

ance. Nevertheless, two regions stand out as having particularly high polymorphism levels: the three coding exons of *Adh* and the two intervening introns.

We now investigate whether estimated θ values for one or more regions are compatible with a particular parametric or average value of θ. For example, noting that the θ estimates for *Adh* Introns 2+3 and the coding region, 0.018 and 0.023, respectively, are substantially larger than for the other regions, taken to be 0.008 per silent site, we would like to know whether the *Adh* regions, ($\hat{\theta}$-0.021) are incompatible with $\theta = 0.008$. Introns 2+3, which is 127 bp long, has seven polymorphic sites. Assuming $\theta = 0.008$ and using (3) we calculate that the probability of seven or more polymorphic sites is 0.066, not significant at the 0.05 level. For the coding region, assuming $\theta = 0.008$ per silent site, the probability of 13 or more polymorphic sites is 0.014, and we conclude that the level of polymorphism in the coding region is too high to be compatible with $\theta = 0.008$. Lumping the *Adh* coding regions and introns 2+3, the probability of 20 or more polymorphisms, with $\theta = 0.008$ per silent site, is 0.013. These calculations assume no recombination within the regions; the probabilities of so many polymorphic sites in these two regions would be even smaller if recombination was taken into account.

An alternative approach to testing for heterogeneity is to calculate a goodness-of-fit statistic. As we will show, the test of goodness-of-fit depends strongly on what assumptions one makes about the levels of recombination within and between the regions being examined. A test statistic suggested recently by one of us (RH) for k regions is

$$X_C^2 = \sum_{i=1}^{k} \frac{(s(i)_{obs} - s(i)_{exp})^2}{\mathrm{Var}(s(i)_{exp})} \qquad (4)$$

where $s(i)_{obs}$ is the observed number of polymorphic sites in the ith region. $s(i)_{exp}$ is the expected number of polymorphic sites in the ith region, given by

$$s(i)_{exp} = \hat{\theta} m_i a_{n-1},$$

where m_i is the length of the ith region, and $\hat{\theta}$ is obtained from (1) using the total number of polymorphic sites and the total length of the k regions in a sample of n genes. And finally the denominator is calculated as

$$\mathrm{Var}(s(i)_{exp}) = \hat{\theta} m_i a_{n-1} + (\hat{\theta} m_i)^2 \sum_{j=1}^{n-1} \frac{1}{j^2}.$$

Under the neutral infinite-sites model, X_c^2 is expected to be approximately χ^2 distributed with $k - 1$ degrees of freedom, when there is no recombination within the regions, free recombination between the regions and n is sufficiently large. Using the critical values from the χ^2 distribution with $k - 1$ degrees of freedom will be conservative when there is some recombination within regions and/or some linkage between the regions. For our case, with very tight linkage between regions and some recombination within regions, this statistic will be very conservative, and simulation results shown in Table 5 show that for five contiguous

TABLE 5

Effect of recombination on 0.05 critical values of X_C^2 and X_L^2

R	X_C^2	X_L^2
0.0	2.5	9.00
0.5	3.62	14.75
1.5	4.75	19.75
2.5	5.25	21.75
3.75	5.5	22.25
5.0	5.5	22.5
7.5	5.25	22.0
10.0	5.25	21.25
20.0	4.75	18.75
50.0	3.75	15.25

Simulation is for five contiguous regions and several values of $R = 2Nr_t$. Samples were generated using a coalescent method described by HUDSON (1983). The method assumes an infinite site neutral model. θ for each of the five regions was 5.6 (which results in an average total number of segregating sites of 82, close to the observed number, 84). The estimated critical values were obtained by generating 10,000 samples of size eleven for each value of R.

regions, the 0.05 critical values are always well below 9.48, the critical value of the χ^2 distribution with 4 degrees of freedom. When this statistic is significantly large, as judged by the critical values of the χ^2 distribution, one can be confident that there is significant heterogeneity, under any assumption about recombination.

Another useful goodness-of-fit statistic, X_L^2, is obtained by replacing $\mathrm{Var}(s(i))_{\mathrm{exp}}$ by $s(i)_{\mathrm{exp}}$ in (4). For k completely linked regions with no recombination within the regions, or for completely free recombination throughout the regions, this statistic is approximately χ^2 distributed with $k - 1$ degrees of freedom under the same neutral model. For intermediate levels of recombination, the critical values of this statistic will be larger than predicted from the χ^2 distribution. As shown in Table 5, the critical value depends on $2Nr_t$, where r_t is the total recombination rate between the ends of the region being considered. When $2Nr_t$ is 5.0, the 0.05 critical value of X_L^2 is approximately 22.5—very much larger than 9.48, the critical value of the χ^2 distribution with four degrees of freedom. For larger or smaller recombination rates the critical values of X_L^2 are smaller than 22.5 as shown in Table 5. If this statistic is not significantly large using the critical values from the χ^2 distribution with k − 1 degrees of freedom, then there is no evidence for heterogeneity among the regions, and no alternative assumptions about recombination will change that conclusion.

The X^2(C and L) test values for the three different subdivisions of the data are presented in Table 6. Dividing the polymorphism data into 3 regions—5′ flanking, *Adh* and *Adh-dup*—or into five intervals containing the same number of silent sites yield no evidence for heterogeneity. Dividing the DNA into five regions with the protein coding and noncoding regions of the two loci separated yields some evidence of heterogeneity, but only if specific assumptions are made about recombination rates. Using the χ^2 distribution assumption, X_L^2 gives a highly significant value, but Table 5 shows that if $2Nr_t$ is around 5.0, the observed value of X_L^2, 21.77, can be obtained with probability slightly higher than 0.05. To obtain a probability of 0.05 or less requires that $2Nr_t$ be greater than 10 or less than 2.5. X_c^2 is not significant when critical values are used from the χ^2 distribution with four degrees of freedom as would be appropriate if the five regions were unlinked to each other and there was no recombination within regions. Simulation results in Table 5 for five contiguous intervals show that there is no level of recombination for which the observed value of X_c^2 2.38, is significant at the 0.05 level.

The goodness of fit tests, therefore, fail to reveal strong evidence of between-region heterogeneity, although there is a suggestion that the *Adh* region, with 20 segregating sites in 319 silent sites, has a level of polymorphism which is incompatible with the other regions. This difference is reflected in the estimated θ's—0.018 and 0.023 for the protein coding regions and small introns of *Adh* compared to values ranging from 0.007 to 0.009 for all other regions. The magnitude of this difference can also be gauged by comparing the observed *Adh* coding and intron θ values with upper and lower 95% confidence limits shown in Figure 3. But even though the *Adh* coding region is more polymorphic than either of its flanking regions, the simple goodness of fit tests do not allow us to reject the null hypothesis. The combination of large evolutionary sampling variances, our uncertainty about recombination rates and the problem of how to subdivide the data makes this approach a tenuous one for detecting selection.

Tests of neutrality based on within- *vs.* between-species comparisons: Even if the goodness of fit test allowed us to reject the completely neutral model, it might be argued that this model of molecular evolution is unrealistically simple. We now consider a more realistic formulation of the neutral model that relaxes an assumption about which sites are selectively constrained. Under this model (see KIMURA 1983) a fraction of all mutations are deleterious and are eliminated by natural selection, the remaining mutations being selectively neutral. Constraint may vary from site to site and from region to region. The model makes no assumptions about the relationship between functionality and constraint. So, the fraction of deleterious mutations can differ among regions of similar function. This provision also allows codons with twofold, fourfold and sixfold redundancy to evolve at different rates [see RILEY (1989) for evidence that this occurs].

The model makes a simple test prediction: levels of

TABLE 6

Distribution of polymorphic "silent" sites

Region	Location		No. silent sites	S_{obs}	S_{exp}	X^2_C	X^2_L
	Start	End					
3 "Loci"							
5′ flanking	−1242	0	1243	30	31.1	0.0067	0.0412
Adh	1	1858	1267.2	34	31.7	0.0211	0.1611
Adh-dup	1859	3275	723.8	17	18.1	0.0184	0.0703
Total						0.05	0.27
5 Regions							
5′ flanking	−1242	0	1243	30	31.1	0.0067	0.0412
Adh 5′	1	777	773	12	19.4	0.1425	2.7985
Adh coding	778	1680	319.2	20	8.0	1.5622	18.0273
Adh 3′	1681	1982	299	5	7.5	0.6509	0.8272
Adh-dup coding	1983	3275	599.8	14	15.0	0.0212	0.0696
Total						2.38	21.77
5 Intervals							
1	−1242	−597	647	14	16.2	0.0980	0.2988
2	−596	50	647	16	16.2	0.0006	0.0025
3	51	698	647	11	16.2	0.8229	1.6691
4	699	1934	647	25	16.2	0.5615	4.7802
5	1935	3275	647	15	16.2	0.0259	0.0889
Total						1.41	6.83

To test for heterogeneity data are clustered three different ways: 3 Loci–5′ flanking–nontranscribed sequence; *Adh*–complete locus (including introns); *Adh-dup*–intergenic region between *Adh* and *Adh-dup* and the complete presumptive coding regions (including introns). 5 Regions–The translated regions of *Adh* and *Adh-dup* (including introns) are segregated from flanking nontranslated regions. 5 Intervals–Equal number of sites in each of 5 intervals. S_{exp} is calculated assuming θ[per site] = 0.0086.

polymorphism within species and levels of divergence between species should be correlated. For example the higher level of silent-site polymorphism in the *Adh* coding region compared to either flanking region implies a reduced level of constraint, and a neutral test prediction is that the *Adh* coding region will have a relatively greater between-species sequence divergence. The HKA test (HUDSON, KREITMAN and AGUADÉ 1987; also KREITMAN and AGUADÉ 1986b) asks whether between-species divergence and within-species polymorphism correspond to the predicted relationship based on a neutral model with selective constraint. The HKA test statistic for the data is

$$X^2 = \sum_{i=1}^{k} \frac{(s_i - \hat{E}(s_i))^2}{\hat{\mathrm{V}}\mathrm{ar}(s_i)} + \sum_{i=1}^{k} \frac{(D_i - \hat{E}(D_i))^2}{\hat{\mathrm{V}}\mathrm{ar}(D_i)}$$

where k is the number of regions being compared, s_i and D_i are the number of polymorphic and diverged sites in region i, respectively, and $\hat{E}(\ )$ and $\hat{\mathrm{V}}\mathrm{ar}(\ )$ are estimates of the expection and variance, respectively. The expected values are formulated from estimated values of a neutral parameter θ_i for each region and a time since divergence of the species. The test statistic is approximately χ^2 distributed with $k - 1$ degrees of freedom. With polymorphism data from only one species, the test assumes the population size of the common ancestor to be the same as the species from which the polymorphism estimate derives.

We have chosen one sequence from *D. melanogaster* (Af^s) and one sequence from a sibling species, *D. simulans* (In: Drsadha) for a between-species comparison. Silent nucleotide differences between the two sequences indicated in Figure 9 are listed by sequence position in Figure 10. A summary of silent and replacement differences between species by region is also presented in Figure 1.

HKA test results for silent sites are presented in Table 7 for three pairs of regions: 5′ Flanking vs. *Adh*, 5′ Flanking vs. *Adh-dup* and *Adh* vs. *Adh-dup*. Two tests were performed for each comparison. The HKA test, as originally formulated, was based on the number of segregating sites in a sample. This test is listed as "Seg" in Table 7. We also include the results from a modified test, listed as "Pwd" in Table 5, which substitutes average pairwise difference (Pwd) for number of segregating sites. Pwd is defined as the average of the number of nucleotide differences, *Sij*, between the ith and jth allele in a sample of n alleles,

$$\mathrm{Pwd} = \frac{\sum_{i,j=1;j<i}^{n} s_{ij}}{n(n-1)/2}.$$

The expectation of Pwd for region i is θ_i; the formula for the variance of the expected pairwise difference is given in TAJIMA (1983). This test may be more sensitive to among-region differences when the observed polymorphism frequency distributions differ across regions. But since the expected variance of Pwd exceeds the expected variance of number of segregating sites, Pwd has less power to reject the null hypothesis

TABLE 7

HKA tests for silent site differences in three regions

Region	Location		Test[a]	Within sp.		Between sp.		$\hat{\theta}$	T(ime)	X^2	P
	Start	End		Obs.	Exp.	Obs.	Exp.				
5′ Flanking	−1242	0	Seg	30	37.5	78	57.7	12.8			
vs. Adh	778	1680		20	12.5	16	19.2	4.3	4.5	3.7	0.054
5′ Flanking	−1242	0	Pwd	11.4	15.1	78	59.3	15.0			
vs. Adh	778	1680		7.6	4.0	16	15.7	4.0	3.9	3.8	0.05
5′ Flanking	−1242	0	Seg	30	27.2	78	71.2	9.3			
vs. Adh-dup	1983	3275		13	15.8	50	41.7	5.4	7.7	1.4	0.25
5′ Flanking	−1242	0	Pwd	11.4	9.2	78	70.1	9.2			
vs. Adh-dup	1983	3275		3.4	5.5	50	42.3	5.5	7.7	1.8	0.18
Adh	778	1680	Seg	20	12.0	16	19.9	4.1			
vs. Adh-dup	1983	3275		13	21.0	50	34.8	7.2	4.9	5.4	0.02
Adh	778	1680	Pwd	7.6	3.4	16	16.9	3.4			
vs. Adh-dup	1983	3275		3.4	7.6	50	38.1	7.6	5.0	6.7	0.01

[a] Seg, number of segregating sites; Pwd, average pairwise number of differences.

when the frequency distributions are the same.

The HKA test is significant at the 0.05 level for the 5′ flanking region vs. the *Adh* coding region, confirming previous results based on four-cutter RFLP data. The patterns of polymorphism and divergence are inconsistent with the neutral model. In order to further distinguish whether the 5′ flanking region or the *Adh* locus (or possibly both regions) is evolving nonneutrally, HKA tests were performed by comparing each of these regions with a third region, the *Adh-dup* locus. Only the *Adh* vs. *Adh-dup* test is significant at the 0.05 level ($P < 0.01$). This result implicates *Adh* and not the 5′ flanking region as the cause of the departure from neutrality.

Test of neutrality based on frequencies at polymorphic sites: The distributions of the observed nucleotide frequencies at polymorphic sites for the 5′ flanking region, the *Adh* locus and the *Adh-dup* locus are presented in Figure 4 along with the predicted frequency distributions under a model of selective neutrality (see TAJIMA 1989). Of particular interest is the depresssed frequency spectrum in the *Adh-dup* region. Considering only silent polymorphisms, seventeen of the 18 sites in this region have frequencies of 1/11 or 2/11 for the less frequent allele. This contrasts with the 5′ flanking region and the *Adh* locus, which have silent polymorphism frequencies 1/11 or 2/11 in 11 of 30 sites and 20 of 34 sites, respectively.

Is the depressed frequency spectrum in the *Adh-dup* region compatible with the neutral model? Tajima has produced a test of the neutral theory prediction about polymorphism frequencies (TAJIMA 1989). Both the number of segregating sites and the average pairwise difference in a sample are unbiased estimates of the neutral parameter but only the average pairwise difference incorporates actual frequencies. Tajima's test

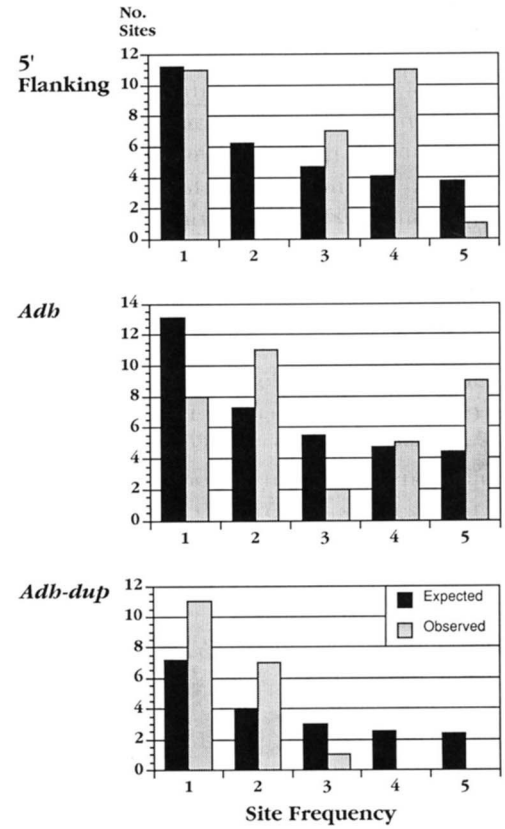


FIGURE 4.—Polymorphism frequency spectra for the 5′ flanking, *Adh* and *Adh-dup* regions along with the predicted distribution under an infinite alleles model (as described in TAJIMA, 1989).

estimates the probability of the two estimates being different by the observed amount. The expected difference under neutrality is, or course, zero.

The observed differences for the *Adh* 5′ flanking region, *Adh* and *Adh-dup* are 0.433, −0.08 and −1.22, respectively. None of the values are statistically significant, although there is an indication of too many low-frequency sites in *Adh-dup*.

Seven of the 18 silent polymorphic sites in *Adh-dup* are the result of the $Fl1^s$ and Wa^s alleles. Excluding these alleles from the sample, all remaining polymorphisms in *Adh-dup* occur at frequency 1/9. This contrasts with the *Adh* 5′ flanking region and the *Adh* locus where, with the $Fl1^s$ alleles removed, frequency one polymorphisms occur at only 12 of 36 and 7 of 38 sites (including insertion/deletions), respectively. With $Fl1^s$ and Wa^s removed from the sample, the difference in the estimated θ's for *Adh-dup* is −2.11, which is significant at the 0.05 confidence level using the Tajima method.

DISCUSSION

The results confirm previous suggestions that higher than expected levels of silent polymorphism in the coding region of *Adh* is the cause of departure from selective neutrality across the *Adh* region. Were this not the case, significant HKA tests would not be expected for *Adh* vs. both the 5′ flanking region and the 3′ *Adh-dup* locus.

This pattern can be explained by the presence of a single balanced polymorphism in the *Adh* locus, for example the threonine-lysine amino acid replacement difference distinguishing the Adh^f and Adh^s alleles. Levels of neutral polymorphism are expected to be elevated when sufficiently tightly linked to a balanced polymorphism, and the effect can be large enough under certain population parameter values to be detected in random samples (STROBECK 1983; HUDSON and KAPLAN 1988). An intuitive explanation for this effect is that neutral mutations persist longer than would be expected under drift alone when they are tightly linked to a balanced polymorphism. The balanced polymorphism, which is assumed to have existed for a long period of time relative to the expected persistence time of a neutral mutation, extends some of its "persistence" onto tightly linked neutral mutations. In other words, it is a tightly linked region that selection holds in the population and not just the balanced polymorphism.

We now evaluate the adequacy of the balanced polymorphism hypothesis by applying a "sliding window" approach to explore how the patterns of polymorphism vary along the DNA (HUDSON and KAPLAN 1988; HUDSON 1990). A window of width w is "placed" at one end of the DNA sequence and the average pairwise number of nucleotide differences is calculated for sites contained within the window. This value is assigned to the nucleotide site at the center of the window. The actual width of the window is adjusted to always contain w silent sites. We have set w to 100 for the present analysis. This value was chosen by trial-and-error to reveal inter-regional heterogeneity. The window is moved along the length of the sequence and the pairwise differences plotted as a function of position. This is an averaging procedure but it can reveal how levels of variation differ among intervals. One might expect, for example, a peak in the graph to be centered around a site containing a balanced polymorphism. Just such a spike occurs in the *Adh* locus and it is centered on position 1490, the site of the threonine-lysine substitution, as indicated in Figure 5.

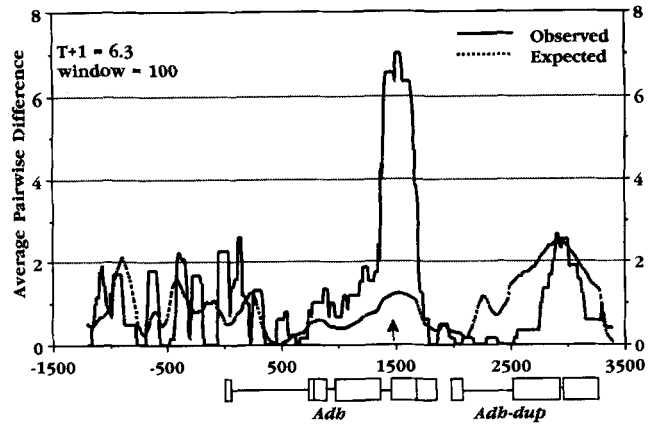


FIGURE 5.—Sliding window for Adh^f/Adh^s comparisons, no-selection model. Predicted values are calculated from between-species data as described in the discussion section. Average pairwise difference is the average number of nucleotide differences between Adh^f and Adh^s alleles. $T + 1$: estimated coalescent time (in units of $2N$ generations) for one *D. melanogaster* allele (Af^s) and one *D. simulans* allele (EMBL: Dsadh01); window size is set to 100 silent base pairs. Arrow at 1490 marks position of the Adh^f/Adh^s protein polymorphism.

A balanced polymorphism is expected to affect levels of neutral linked polymorphism but not the corresponding level of neutral sequence divergence. Selective constraint, in contrast, is expected to affect both levels of polymorphism and divergence equally. This is the basis for the HKA test as described above. Comparison of polymorphism levels and appropriately scaled divergence levels between closely related species, both of which can be plotted in a sliding window, can reveal where peaks or troughs of polymorphism do not conform to selective constraint.

The method for using between-species data to estimate mutation parameters is described fully elsewhere (HUDSON 1990). Briefly, each site can have a different neutral parameter, denoted θ_i. Let T be the time since the divergence of the two species, measured in units of $2N$ generations. Then under the neutral theory, for small $\theta_i(T + 1)$ the probability that two sequences at site i are different is approximately $\theta_i(T + 1)$. θ_i is estimated by dividing the number of differences between the two sequences in a window of width w centered at site i by $(T + 1)w$. Assuming constant population size, the estimated neutral parameter value for each site is also the expected average pairwise difference within species. This is because, under the neutral model the expected average pairwise differ-

ence between a sample of genes is an unbiased estimator of the neutral parameter θ.

The time T can be thought of as scaling factor between polymorphism and divergence. The value we use here, $T + 1 = 6.3$, is estimated from all the data, as described in HUDSON, KREITMAN and AGUADÉ (1987). Another possibility is to choose a value of T that gives a good fit between polymorphism and divergence for one subregion.

The "window" method visually represents polymorphism and divergence data at every "position" (actually a window centering on a position) along the length of the DNA. Theoretical predictions can be generated for every site but the fit between observation and prediction must be made by inspection. Therefore, this approach does not lend itself to formal hypothesis testing, but neither does it require *a priori* decisions about how to subdivide and sum the data (other than to set the window size). The method compromises the ability to test hypotheses for the ability to explore the data without regards to scale (subdivision length).

Observed and predicted polymorphism levels are shown in Figure 5. A qualitative summary of the plot is as follows. There is a close correspondence between polymorphism and divergence in the 5′ flanking region and through the 5′ half of *Adh*. For example, *Adh* Intron 1 exhibits a trough of divergence and a corresponding trough of polymorphism. This is not unexpected: the 3′ end of the intron contains the initiation site and transcription signals for larval *Adh* mRNA transcription (FISCHER and MANIATIS 1988).

There is a large excess of observed over predicted polymorphism in *Adh* Intron 3 and Exon 3 and the excess is centered on the threonine-lysine amino acid replacement substitution. This contrasts with the *Adh-dup* locus, where silent sites are evolving on average at approximately twice the rate of *Adh* but which shows a moderate deficiency of observed polymorphism. This deficiency will be analyzed in greater detail later in this section.

First, we can ask how well the assumption of a balanced polymorphism at position 1490 brings the predicted polymorphism level into accord with the observed level. Two additional parameters are required to fit a balanced polymorphism model, a gene frequency for the site under selection and a recombination parameter. Let R denote $2Nr$, where r is the recombination rate per generation between adjacent bases. Following the procedure outlined by HUDSON and KAPLAN (1988) but using the θ_i values for each site (rather than assuming a single parametric value) a good fit is obtained assuming a recombination rate of $R = 0.005$, as shown in Figure 6. The best available estimate of R is 0.012, based on per nucleotide recombination and mutation rate estimates in *D. melanogaster* (CHOVNICK, GELBART and MCCARRON 1977; also see HUDSON and KAPLAN 1988), a factor of only two greater than our "best" fit value. Rather than assuming a single θ value and applying it to all sites (as in HUDSON and KAPLAN 1988), calculating individual θ_i values brings the best-fit R value closer to the *a priori* predicted value by a factor of three. Given the uncertainty of the recombination and mutation rate estimates needed to obtain predicted values, we consider observed and predicted silent polymorphism values to be in excellent agreement.

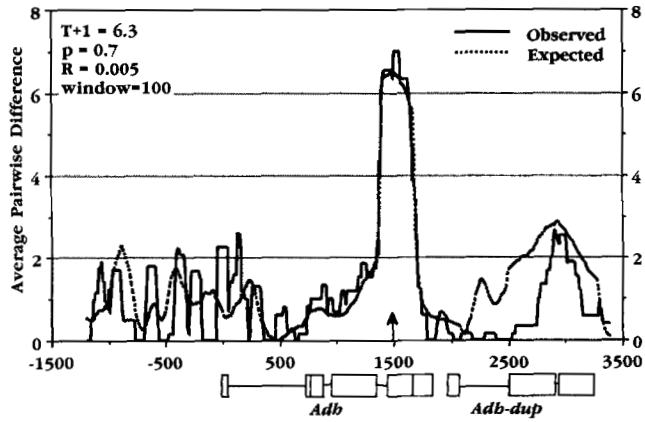


FIGURE 6.—Sliding window for Adh^f/Adh^s comparisons, selection model with balanced polymorphism at position 1490 (indicated by vertical arrow). The frequency, p, of the Adh^s slow allele is set to 0.7; using a value of 0.005 for the recombination parameter, $R = 2Nr$, yields a good fit between observed and expected polymorphism.

However, another prediction of the balanced hypothesis does not fare as well. Under the balance hypothesis the higher than expected neutral polymorphism level is caused by the accumulation of differences between the two allelic classes. A corollary is that levels of silent polymorphism within an allelic class should be slightly reduced compared to the neutral expectation for a random sample of the same size. As shown in Figure 7 this is not the case. A higher than expected level of silent polymorphism persists within the Adh^s allele class. Also, the peak of polymorphism in the *Adh-dup* locus seen in Figures 5 and 6 are entirely represented within the Adh^s class.

The cause of this within-Adh^s polymorphism is a cluster of sites distinguishing the Wa^s and standard Adh^s alleles, as revealed in Tables 2 and 3. What has maintained these two old Adh^s lineages?

We showed, in the previous section, that recombination suppression is unlikely to be maintaining the two distinct haplotypes. One way to resurrect the recombination-suppression hypothesis is if the Wa^s framework has recently recombined away from an allele carrying a recombination-suppressing inversion. This model has the virtue of being able to account for the lack of a Wa^s pattern in the *Adh* 5′ flanking region by specifying the interval within which the recombi-

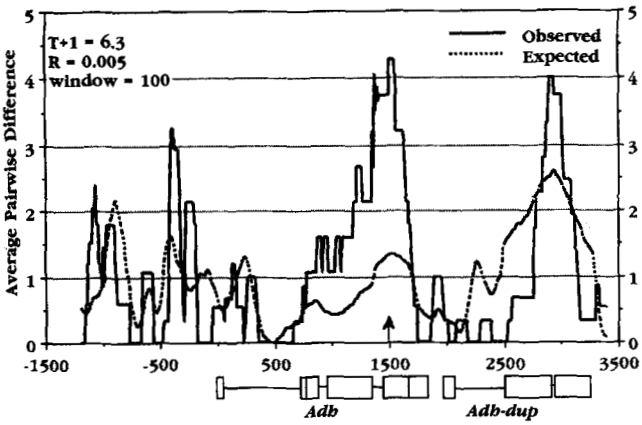


FIGURE 7.—Sliding window for Adh^s only comparisons, no-selection model. The observed average pairwise difference is the average of all pairwise comparisons within Adh^s. The greater than expected variation in *Adh* (and also the spike of polymorphism in *Adh-dup*) is caused by 19 differences between Wa^s and Fl^{1s} and the remaining alleles.

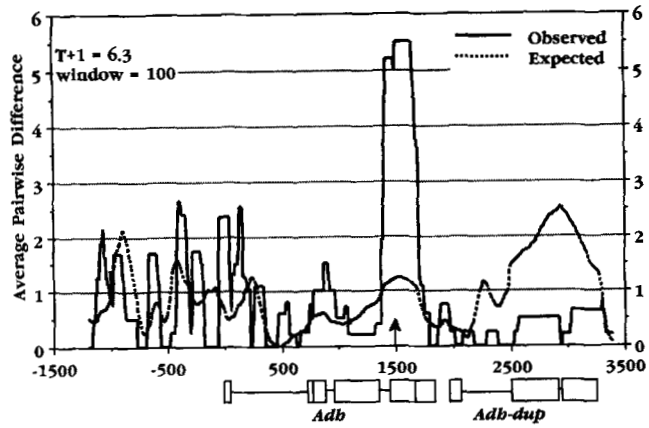


FIGURE 8.—Sliding window Adh^f/Adh^s with Wa^s and Fl^{1s} excluded, no-selection model.

nation (or gene conversion) breakpoint(s) occurred. One possible candidate is the inversion *In(2L)t*. This naturally occurring inversion polymorphism in *D. melanogaster* is widely distributed in low-latitude populations but is rare in North America (ASHBURNER and LEMEUNIER 1976; VOELKER *et al.* 1978). The proximal breakpoint of *In(2L)t*, cytologically positioned at 34A8-9 (LINDSLEY and GRELL 1968), is very close to *Adh* (33D1-34E5) and the allozyme polymorphism is known to be in strong linkage equilibrium with the inversion (KOJIMA, GILLESPIE and TOBARI 1970; MUKAI, METTLER and CHIGUSA 1971; LANGLEY, TOBARI and KOJIMA 1974; METTLER, VOELKER and MUKAI 1977; VOELKER *et al.* 1978; KNIBB 1982; VAN DELDEN and KAMPING 1989).

Restriction fragment length polymorphism in the *Adh* locus has recently been investigated in 40 wild-derived *In(2L)t* alleles from a Spanish population using a four-cutter analysis (AGUADÉ 1988). Interestingly, the Wa^s framework of five distinguishing sites was represented three times. In addition, 18 lines representing five Wa^s partial-haplotypes were also present. This is distinctly different from the North American samples, where partial-Wa^s alleles are absent. Unfortunately, of the 39 standard gene arrangement alleles included in the Spanish population study, only three carried the Adh^s allele. Therefore, the frequency of the Wa^s haplotype in standard gene arrangement Adh^s alleles is not known. This hypothesized connection between *In(2L)t* and the Wa^s haplotype deserves further investigation.

The presence of so many fixed nucleotide differences between Wa^s and standard Adh^s suggests that these alleles have been maintained in the species as distinct entities for a considerable time. How likely is their presence under neutrality? Unfortunately, it is difficult to make this assessment without a good estimate of the per nucleotide recombination rate. It seems to us unlikely that natural selection has held the Wa^s mutations in linkage disequilibrium because all the mutations are silent and a majority are in introns. If the presence of the two old lineages is a historical artifact, as it might be if the species was previously subdivided into two semi-isolated stocks, then similar divergences should be expected in other parts of the genome. Additional sequence data may eventually allow us to statistically resolve whether or not Wa^s has been selectively maintained.

Having considered possible "causes" of within-Adh^s polymorphism, does a higher than expected level of polymorphism remain in the Adh^f and Adh^s alleles when the Wa^s and $Fl1^s$ alleles are removed? Under the balanced polymorphism hypothesis, an excess across all alleles should still be present in the reduced sample. The resulting pattern of polymorphism, given in Figure 8, shows a clear excess of silent polymorphism around position 1490. The effect of removing Wa^s and $Fl1^s$ is a small reduction in the height of the peak from an average pairwise difference of 6.9 to 5.5. The expected level remains at approximately 1.5. The balanced polymorphism hypothesis between Adh^f and Adh^s cannot be rejected even after accounting for the higher-than-expected silent polymorphism within the Adh^s class, and the "excess" polymorphism is only slightly less dramatic.

One final comment should be made about our choice of the Adh^f-Adh^s substitution at position 1490 as being the site under balancing selection. We have no direct evidence to support the choice of this site. Rather it comes from other kinds of evidence about the allozyme polymorphism (see VAN DELDEN 1982; reviewed in LEWONTIN 1985). Our data indicate only the presence of a balanced polymorphism at a site within a small region encompassing position 1490. Another tightly linked site may be the site of selection. It is even possible that a combination of tightly linked sites is being selected. We cannot distinguish between these hypotheses with the current data.

The removal of Wa^s and $Fl1^s$ alleles has a dramatic effect on levels of polymorphism in the *Adh-dup* locus, as shown in Figure 8. Now the silent polymorphism levels are dramatically lower than the expected level. Although the evidence is weak, the lower than expected level of silent variation as well as the depressed frequency spectrum may be indications of a recent selective substitution in the *Adh-dup* locus or at a site further downstream. Both effects are predicted for neutral mutations linked to a selectively favored substitution (MAYNARD SMITH and HAIGH 1974; KAPLAN, HUDSON and LANGLEY 1989). Additional data will be needed to evaluate this hypothesis.

In summary, we can state with some certainty that the patterns of polymorphism in *D. melanogaster* encompassing the *Adh* and *Adh-dup* loci do not conform to a model of neutral molecular evolution, and selective constraints do not account for the observed patterns. A proposed balanced polymorphism in *Adh* is consistent with many aspects of the data. The protein polymorphism seems a convenient candidate as it is located within a peak of higher than expected silent polymorphism; it also conforms to other kinds of evidence for the allozyme polymorphism being a balanced polymorphism. Support for a selective "sweep" in *Adh-dup* is weak but is consistent with certain features of the data. This hypothesis deserves further attention. And finally the Wa^s polymorphisms in *Adh* and *Adh-dup* remain largely unexplained but may not be maintained by selection. It is possibly an artifact of the particular evolutionary history of this region–such as a previous linkage to an inversion–or may reflect a previous geographic split within the species.

Whether or not multiple causes of polymorphism can be disentangled one from another remains a substantial challenge. For example, is it statistically justifiable to remove some of the sample data in order to evaluate whether the residual data requires additional explanation? Essentially this is what we have done in order to identify multiple causes. It would also be desirable to have tests of evolutionary hypotheses that do not require *a priori* subdivision of the data. If the *Adh* and *Adh-dup* region does in fact have a complex evolutionary history involving multiple selective forces and historical effects, then whether or not this is typical of the *Drosophila* genome becomes a very important question.

We thank D. DENKER for help with the recombination suppression experiment, A. BERRY for suggesting the use of upper and lower confidence limits for θ, and M. TAYLOR and J. MCDONALD for critical comments on the manuscript. The work is supported by National Institutes of Health (NIH) grant GM39355 and National Science Foundation grant DCB-8646262 to M.K. and NIH grant GM42447 to R.H.

LITERATURE CITED

AGUADÉ, M., 1988 Restriction map variation at the *Adh* locus of *Drosophila melanogaster* in inverted and noninverted chromosomes. Genetics **119:** 135–140.

AGUADÉ, 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.

AGUADÉ, 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.

AQUADRO, C. F., 1989 Contrasting levels of DNA sequence variation in *Drosophila* species revealed by "six-cutter" restriction map surveys in *Molecular Evolution,* UCLA Symposium on Molecular and Cellular Biology, New Series, Vol. 122, edited by M. CLEGG and S. O'BRIEN. Alan R. Liss, New York.

AQUADRO, C. F., K. M. LADO and W. A. NOON, 1988 The *Rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875–888.

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

ASHBURNER, M., and F. LEMEUNIER, 1976 Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*Sophophora*). I. Inversion polymorphisms in *D. melanogaster* and *D. simulans*. Proc. R. Soc. Lond. **193:** 137–157.

BEECH, R. N., and A. J. LEIGH BROWN, 1989 Insertion-deletion variation at the *yellow-achaete-scute* region in two natural populations of *Drosophila melanogaster*. Genet. Res. **53:** 7–15.

CHOVNICK, A., W. GELBART and M. MCCARRON, 1977 Organization of the *Rosy* locus in *Drosophila melanogaster*. Cell **11:** 1–10.

COHN, V. H., and G. P. MOORE, 1988 Organization and evolution of the alcohol dehydrogenase gene in *Drosophila*. Mol. Biol. Evol. **5:** 154–166.

COLLET, C., 1988 Recent origin for a thermostable alcohol dehydrogenase allele of *Drosophila melanogaster*. J. Mol. Evol. **27:** 142–146.

COYNE, J. A., and M. KREITMAN, 1986 Evolutionary genetics of two sibling species, *Drosophila simulans* and *Drosophila sechellia*. Evolution **40:** 673–691.

DAVÍD, J. R., and P. CAPY, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

DENKER, A. D., 1990 Gene structure at the *Adh-dup* locus in *Drosophila melanogaster*. Undergraduate thesis, Princeton University.

EANES, W. F., J. LABATE and J. W. AJIOKA, 1989 Restriction-map variation within the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 492–502.

EANES, W. F., J. W. AJIOKA, J. HEY and C. WESLEY, 1989 Restriction-map variation associated with the G6PD polymorphism in natural populations. Mol. Biol. Evol. **6:** 384–397.

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

FAGLES, N., 1989 Structure of *Adh-dup*, an ancient tandem duplication of *Adh* in *Drosophila melanogaster*. Undergraduate thesis, Princeton University.

FISCHER, J. A., and T. MANIATIS, 1988 Drosophila *Adh*: a promoter element expands the tissue specificity of an enhancer. Cell **53:** 451–461.

GOLDING, G. B., C. F. AQUADRO and C. H. LANGLEY, 1986 Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA **83:** 427–431.

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

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7.**

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process

in models with selection and recombination. Genetics **120:** 831–840.

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

KAPLAN, N. L., HUDSON, R. R. and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KNIBB, W. R., 1982 Chromosome inversion polymorphisms in *Drosophila melanogaster*. II. Geographic clines and climatic associations in Australasia, North America and Asia. Genetica **58:** 213–223.

KOJIMA, K., J. GILLESPIE and Y. N. TOBARI, 1970 A profile of Drosophila species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities and linkage disequilibrium in glucose-metabolizing systems and some other enzymes. Biochem. Genet. **4:** 627–637.

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

KREITMAN, M., 1990 Detecting selection at the level of DNA, in *Evolution at the molecular level*, edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM, Sinauer Associates, Sunderland, Mass.

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

KREITMAN, M., and M. AGUADÉ, 1986b Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. Genetics **114:** 93–110.

KREITMAN, M., and L. L. LANDWEBER, 1989 A strategy for producing single-stranded DNA in the polymerase chain reaction: a direct method for genomic sequencing. Gene Anal. Tech. **6:** 84–88.

LANGLEY, C. H., E. MONTGOMERY and W. F. QUATTLEBAUM, 1983 Restriction map variation in the *Adh* region of *Drosophila*. Proc. Natl. Acad. Sci. USA, **79:** 5631–5635.

LANGLEY, C. H., Y. N. TOBARI and K. KOJIMA, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics **78:** 921–936.

LANGLEY, C. H., A. E. SHRIMPTON, T. YAMAZAKI, N. MIYASHITA, Y. MATSUO and C. F. AQUADRO, 1988 Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. Genetics **119:** 619–629.

LAURIE-AHLBERG, C. C., and L. F. STAM, 1987 Use of *P*-element mediated transformation to identify the molecular basis of naturally occurring variants affecting *Adh* expression in *Drosophila melanogaster*. Genetics **115:** 129–140.

LAURIE, C. C., and L. F. STAM, 1988 Quantitative analysis of RNA produced by Slow and Fast alleles of *Adh* in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA, **85:** 5161–5165.

LEIGH BROWN, A. J., 1983 Variation at the 87A heat shock locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA, **80:** 5350–5354.

LEMEUNIER, F., J. R. DAVID, L. TSACAS and M. ASHBURNER, 1986 The *melanogaster* species group, pp. 147–265 in *The Genetics and Biology of Drosophila*, Vol. 3e, edited by M. ASHBURNER, H. L. CARSON and J. N. THOMPSON, JR. Academic Press, London.

LEWONTIN, R. C., 1985 Population genetics. Annu. Rev. Genet. **19:** 81–102.

LINDSLEY, D. L., and E. H. GRELL, 1968 *Genetic Variations of Drosophila melanogaster*. Carnegie Inst. Wash. Publ. 627.

MACPHERSON, J. N., B. S. WEIR and A. J. LEIGH BROWN, 1990 Extensive linkage disequilibrium in the *achaete-scute* complex of *Drosophila melanogaster*. Genetics **126:** 121–129

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favorable gene. Genet. Res. **23:** 23–35.

METTLER, L. E., R. A. VOELKER and T. MUKAI, 1977 Inversion clines in populations of *Drosophila melanogaster*. Genetics **87:** 169–176.

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

MUKAI, T., L. E. METTLER and S. I. CHIGUSA, 1971 Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA, **68:** 1065–1069.

RILEY, M. A., 1989 Nucleotide sequence of the *Xdh* region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol. Biol. Evol. **6:** 33–52.

RILEY, M. A., M. E. HALLAS and R. C. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the *Xdh* locus in *Drosophila pseudoobscura*. Genetics **123:** 359–369.

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., C. F. AQUADRO and C. H. LANGLEY, 1988 Restriction-map variation in the *Notch* region of *Drosophila melanogaster*. Mol. Biol. Evol. **5:** 30–40.

SIMMONS, G. M., M. KREITMAN, W. F. QUATTLEBAUM and N. MIYASHITA, 1989 Molecular analysis of the alleles of alcohol dehydrogenase along a cline in *Drosophila melanogaster*. I. Maine, North Carolina and Florida. Evolution **43:** 393–409.

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

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

VAN DELDEN, W., 1982 The alcohol dehydrogenase polymorphism in *Drosophila melanogaster*. Selection at an enzyme locus. Evol. Biol. **15:** 187–222.

VAN DELDEN, W., and W. A. KAMPING, 1989 The association between the polymorphisms at the *Adh* and *αGpdh* loci and the *In(2L)t* inversion in *Drosophila melanogaster* in relation to temperature. Evolution **43:** 775–793.

VOELKER, R. A., C. C. COCKERHAM, F. M. JOHNSON, H. E. SCHAFFER, T. MUKAI and L. E. METTLER, 1978 Inversions fail to account for allozyme clines. Genetics **88:** 515–527.

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

WOODRUFF, R. C., and M. ASHBURNER, 1979 The genetics of a small autosomal region of *Drosophila melanogaster* including the structural gene for alcohol dehydrogenase. II. Lethal mutations in the region. Genetics **92:** 133–149.

Communicating editor: A. G. CLARK

APPENDIX

Polymorphisms of 11 genes of *D. melanogaster* are given in Table 8. The DNA sequence of *D. melanogaster* (Af^s) allele is shown in Figure 9; Figure 10 shows positions containing silent substitutions between *D. melanogaster* (Af^s) and *D. simulans* (Cohn).

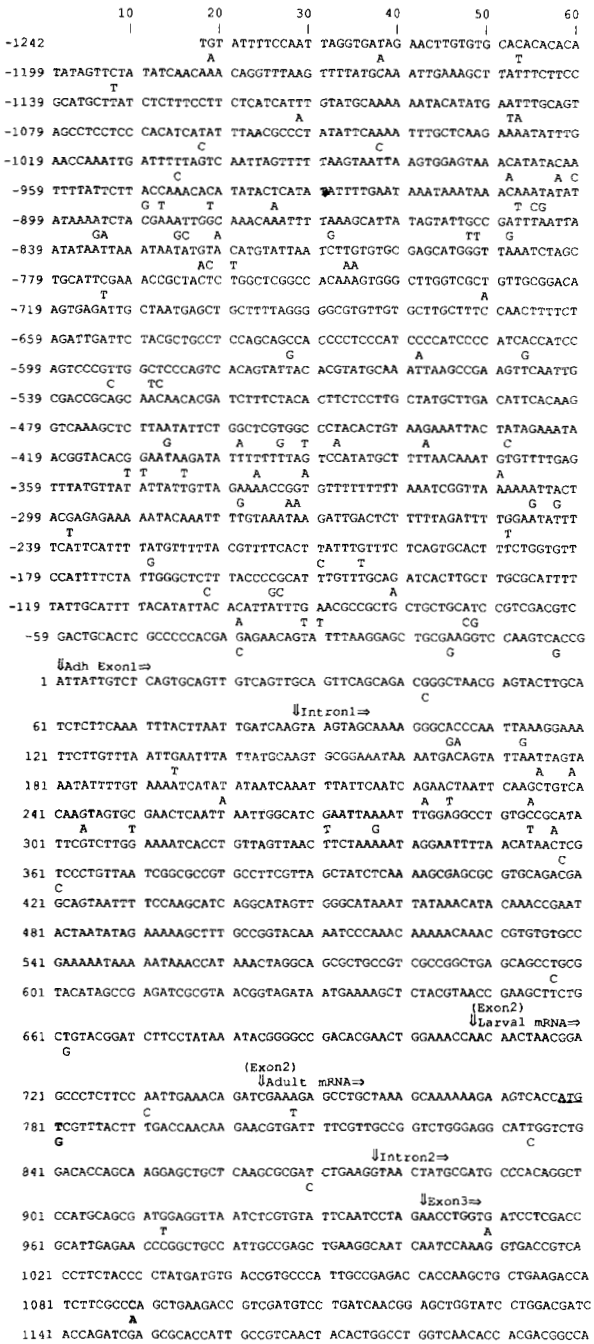
TABLE 8

Polymorphisms in a sample of eleven genes of *D. melanogaster*

Site No.[a]	Position in *Af* 5′ Flanking, *Adh*, *Adh-dup*	Ancestral sequence	Mutated to	Comments
5′ Flanking				
1	−1111	A	T	
2	−1105:6		TATG	Insertion
3	−1092	A or G	G or A	
4	−1068	A	T	
5	−1042	C	A	
6	−959:60	A		Deletion
7	−1017	C	G	
8	−949	G	T or A	A/T Poly.
9	−927:8		CTCATATA	Insertion
10	−923	G OR A	A OR G	
11	−921	A OR T	T OR A	
12	−919	A	T	
13	−897:8		AT	Insertion
14	−822	A	G	
15	−635	C	T	
16	−624:5		TCCATC	Insertion
17	−623	C	T	
18	−618	A	C	
19	−613	C	T	
20	−432	T	A	
21	−388	C	T	
22	−384	A	G	
23	−382	G	T	
24	−374	C	T	
25	−369	A	G	
26	−362	G OR A	A OR G	
27	−359	T	A	
28	−351	A	T	
29	−320:2		TTT, TA,AA,TAA	Complex insertion
30	−255	A OR C	C OR A	
31	−242	T	G	
32	−230	T	A	
33	−227	T	A	
34	−2	C	T	
35	−1	C	G	
36	0	G	C	
Adh intron 1				
1	107	C	A	
2	113	G	A	
3	143	A	G	
4	169	T	G	
5	173	A	G	
6	175	A	T	
7	287	G	T	
8	293	G	T	
9	304	G	C	
10	447:8–476:7	GTTGGGCATAA ATTATAAACAT ACAAACC	TAATATACTAA TACTAATACTA ATACTAATATA A	Ancestral sequence replaced
11	516	C	G	
12	550:1		AATAAACCATA AACTAAGGCTG CTCAGCCGGCGA CGGC	Insertion
13	586	G	T	
Adh exon 2				
14	713	C	A	
15	816	T	G	

Site No.[a]	Position in *Af* 5′ Flanking, *Adh*, *Adh-dup*	Ancestral sequence	Mutated to	Comments
Adh intron 2				
16	896	A	G	
17	925	C	T	
Adh exon 3				
18	1068	C	T	
19	1229	T	C	
20	1235	C	A	
21	1283	C	A	
Adh intron 3				
22	1354	C	G	
23	1362	G	A	
24	1388	A	G	
25	1400	T	A	
26	1405	T	A	
Adh exon 4				
27	1425	C	A	
28	1431	T	C	
29	1443	C	G	
30	1452	C	T	
31	1490	A	C	Lys → Thr
32	1518	C	T	
33	1527	C	T	
34	1557	A OR C	C OR A	Poly. between sim./mau.
35	1596	G	A	
Adh 3′ nontranslated				
36	1693	C	A	
37	1698:1708	10–16"A" run		Length poly.
38	1740	C	G	
Adh-dup noncoding?				
1	1908	T	A	
2	1925	G	A	
3	1937	C	T	
Adh-dup intron 1				
4	2080:1		GT	Insertion
5	2130	C	T	
6	2303:11	9–11 "T" RUN		Length poly.
7	2347	T	A	
8	2378:2405		DELETED	
Adh-dup exon 2				
9	2789	G	A	
10	2807	A	T	
Adh-dup intron 2				
11	2921	C	A	
12	2923	A	G	
13	2926	C	T	
14	2943	T	A	
15	2944:5	T		Deletion
16	2950	T	A	
17	2954	C	G	
18	2957	T	G	
Adh-dup exon 3				
19	3110	A	G	
20	3114	G	A	Val → Ile
21	3266	G	T	Stop codon
22	3277	G	A	
Adh-dup 3′ nontranslated?				
23	3307	G	A	
24	3340:80		Element insertion in region	

[a]Site numbers are the same as those given in Tables 1, 2 and 3.



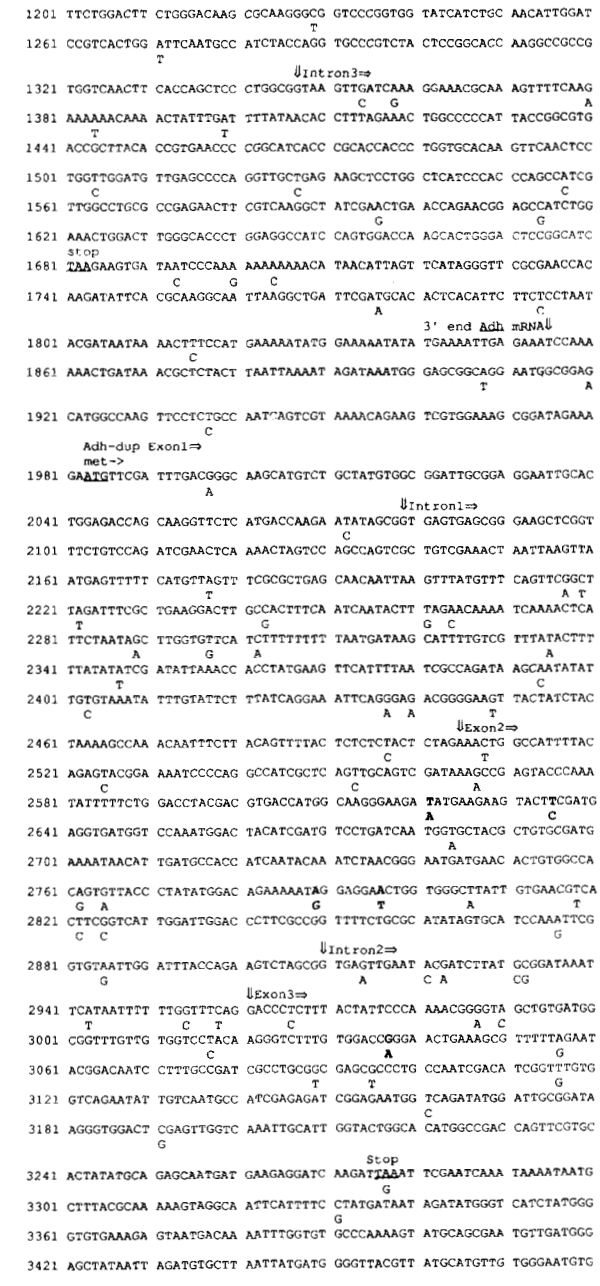


FIGURE 9.—DNA sequences of *D. melanogaster* (Af^s) allele. *Adh* transcription begins at position 1. *Adh-dup* coding region begins at position 1982. Single nucleotide substitution differences between Af^s and *D. simulans* (In: Drsadha) are printed below the sequence. Boldface differences are amino acid replacements.

5' Flanking

-1241 -1222 -1207 -1192 -1111 -1088 -1087 -1062 -1042 -1005 -968
-962 -960 -949 -947 -941 -934 -907 -905 -904 -894 -893 -885 -884
-880 -868 -852 -851 -848 -822 -821 -819 -806 -805 -773 -730 -632
-618 -606 -592 -588 -587 -466 -458 -453 -450 -447 -437 -428 -410
-409 -404 -396 -390 -369 -338 -332 -331 -305 -302 -297 -248 -228
-209 -204 -161 -154 -153 -140 -98 -90 -89 -72 -71 -39 -15 -3

Adh

42 105 106 113 134 175 179 200 222 225 234 244 250 271 277 294 297
358 361 597 662 731 758 834 870 913 950 1229 1271 1354 1358 1380
1384 1399 1504 1527 1557 1596 1614 1683 1690 1694 1776 1794 1815

Adh-dup

1908 1920 1937 2072 2177 2217 2219 2222 2243 2261 2264 2289 2297
2302 2335 2347 2394 2403 2437 2440 2449 2497 2508 2525 2553 2567
2684 2762 2765 2807 2819 2822 2825 2876 2885 2914 2921 2923 2931
2932 2943 2954 2958 2966 2987 2990 3017 3056 3089 3095 3116 3161
3191 3277 3331

FIGURE 10.—Positions containing silent substitutions between *D. melanogaster* (Af^s) and *D. simulans*. (In: Drsadha)