J Mol Evol (1985) 22:289–300

Journal of Molecular Evolution

Phylogenetic Analysis of Polymorphic DNA Sequences at the Adh Locus in *Drosophila melanogaster* and Its Sibling Species

J. Claiborne Stephens and Masatoshi Nei

Center for Demographic and Population Genetics, University of Texas at Houston, PO Box 20334, Houston, Texas 77225, USA

Summary. Recent sequencing of over 2300 nucleotides containing the alcohol dehydrogenase (Adh) locus in each of 11 *Drosophila melanogaster* lines makes it possible to estimate the approximate age of the electrophoretic "fast–slow" polymorphism. Our estimates, based on various possible patterns of evolution, range from 610,000 to 3,500,000 years, with 1,000,000 years as a reasonable point estimate. Furthermore, comparison of these sequences with those of the homologous region of *D. simulans* and *D. mauritiana* allows us to infer the pattern of evolutionary change of the *D. melanogaster* sequences. The integrity of the Adh-f electrophoretic alleles as a single lineage is supported by both unweighted pair-group method (UPGMA) and parsimony analyses. However, considerable divergence among the Adh-s lines seems to have preceded the origin of the Adh-f allele. Comparisons of the sequences of *D. melanogaster* genes with those of *D. simulans* and *D. mauritiana* genes suggest that the split between the latter two species occurred more recently than the divergence of some of the present-day Adh-s genes in *D. melanogaster*. The phylogenetic analyses of the *D. melanogaster* sequences show that the fast–slow distinction is not perfect, and suggest that intragenic recombination or gene conversion occurred in the evolution of this locus. We extended conventional phylogenetic analyses by using a statistical technique for detecting and characterizing recombination events. We show that the pattern of differentiation of DNA sequences in *D. melanogaster* is roughly compatible with the neutral theory of molecular evolution.

Offprint requests to: J.C. Stephens

Key words: Gene genealogy — Phylogenetic inference — Intragenic recombination — Gene conversion — Neutral theory

Introduction

One of the best-studied polymorphisms in population genetics is the electrophoretic "fast" (f) and "slow" (s) variation at the alcohol dehydrogenase (Adh) locus of *Drosophila melanogaster*. This apparently worldwide polymorphism (Johnson and Schaffer 1973; Oakeshott et al. 1982) is caused by a single amino acid change (Thatcher 1980). In a recent study of DNA sequence variation at this locus, Kreitman (1983) demonstrated extensive silent polymorphism within each electromorph. The Adh gene sequences of the related species *D. simulans* and *D. mauritiana* were also studied recently (Bodmer and Ashburner 1984; Cohn et al. 1984). These sequence data can be used for studying the evolutionary relationship of polymorphic genes belonging to the same species as well as to different species.

The purpose of this paper is threefold: (1) to analyze the possible phylogenetic relationships among the available DNA sequences, including the possibility of gene conversion or intragenic recombination; (2) to estimate the age of the fast–slow electrophoretic polymorphism in *D. melanogaster*, and the times of splitting of *D. melanogaster*, *D. simulans*, and *D. mauritiana*; and (3) to examine whether these data are consistent with the neutral theory of molecular evolution. In the course of accomplishing these goals, we point out the pitfalls to which con-

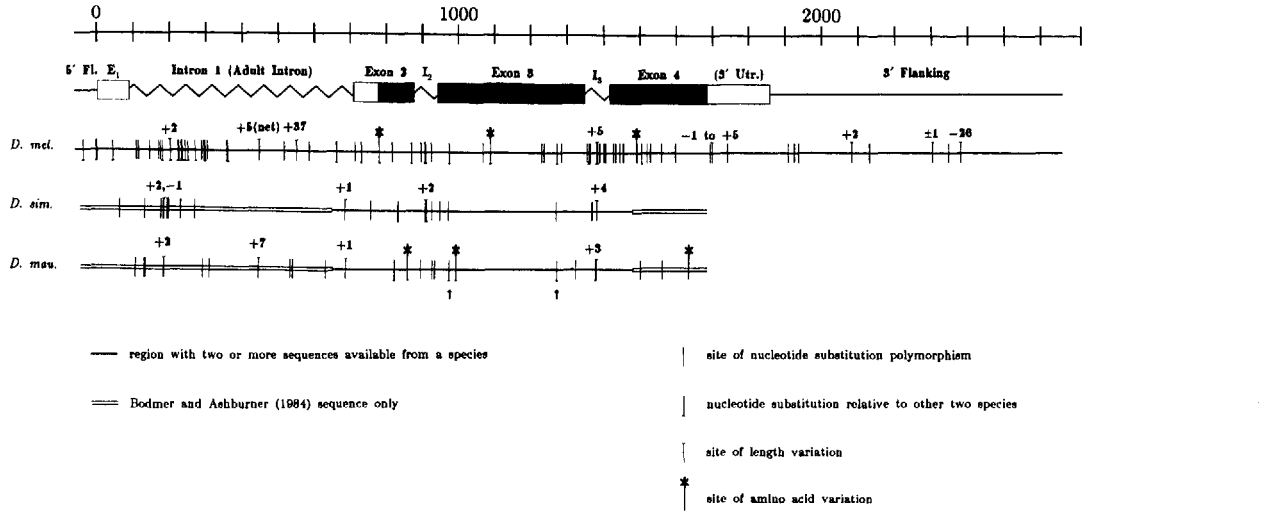

Fig. 1. Physical distribution of all mutational events. The two arrows below *D. mauritiana* sequence indicate polymorphism in *D. simulans*, with fixation of the alternative nucleotides in each of the other two species. Scale is identical to Kreitman's (1983). Several discrepancies in the studies of Bodmer and Ashburner (1984) and Cohn et al. (1984) are resolved: Positions −117, −108, 45, 122, 222, 533–544, and 672 of Cohn et al. no longer show any differences (unresolved or otherwise) from corresponding sequences of Bodmer and Ashburner (V. Cohn, personal communication). A hypervariable region beginning at position 601 of Cohn et al. was realigned to reflect an insertion polymorphism in each of *D. simulans* and *D. mauritiana*, in addition to variability seen by Bodmer and Ashburner. Positions 161, 294, 800, 1184, 1267, and 1348 (only *D. mauritiana*) of Bodmer and Ashburner no longer show any differences (unresolved or otherwise) from corresponding sequences of Cohn et al. or Kreitman (1983) (M. Ashburner, personal communication). Also, correcting Bodmer and Ashburner's sequences, position 238 is G and position 284 is C in *D. simulans* sequence; consensus C at 1005 is really G, with variant C in *D. simulans*; A → G transition shown at 1436 in *D. mauritiana* probably belongs at 1435; position 1583 is A in *D. mauritiana* sequence; and variant A shown in *D. mauritiana* at 1640 should be at 1643. Three pairs of insertion/deletion events, near Bodmer and Ashburner's positions 152, 178, and 225, were realigned to reflect six base substitutions

ventional methods are subject in the presence of recombination, and suggest certain modifications.

Statistical Overview

D. melanogaster

Kreitman (1983) sequenced 11 genes cloned from various *D. melanogaster* lines derived from diverse geographic areas including Washington (Wa-s, Wa-f), Florida (Fl-1s, Fl-2s, Fl-f), France (Fr-s, Fr-f), Africa (Af-s, Af-f), and Japan (Ja-s, Ja-f) (Table 1). Each sequence was at least 2379 base pairs (bp) long, including the Adh coding region, introns, adult and larval mRNA leader sequences, and both 5′ and 3′ flanking regions (Fig. 1). Forty-three segregating nucleotide sites and six DNA length polymorphisms were detected. None of the variable nucleotide sites had more than one variant, although two of the length variations involved homonucleotide repeats (i.e., AA . . . A and TT . . . T). The physical distribution of all polymorphisms is shown in Fig. 1, and the distribution of the *D. melanogaster* variations among all lineages is shown in Table 1. Almost half of the segregating nucleotide sites in *D. melanogaster* are located within a stretch of less than 600 bp centered on the polymorphic position responsible for the single amino acid change. This region extends from the 3′ end of exon 3 to the 3′ untranslated portion of exon 4 (Fig. 1).

The numbers of nucleotide differences (n_d) among the 11 *D. melanogaster* sequences, excluding insertions and deletions, are shown in Table 2 above the diagonal. This table also includes (below the diagonal) the numbers of nucleotide substitutions per site (d), estimated by

$$d = -\tfrac{3}{4} \log_e(1 - 4p/3) \qquad (1)$$

(Jukes and Cantor 1969), where $p = n_d/n$. Here, n is the number of nucleotides compared (in the present case, n = 2379). The variance of d is given by

$$V(d) = p(1 - p)/n(1 - 4p/3)^2 \qquad (2)$$

(Kimura and Ohta 1972). The d value varies considerably with sequence pair, the largest value being d = 0.0123 between Wa-s and Ja-f. The average value of d is 0.0066 (Table 3).

D. simulans and D. mauritiana

Much (1776 bp) of the region sequenced by Kreitman (1983) was recently studied by Bodmer and Ashburner (1984) for *D. simulans*, *D. mauritiana*, and *D. orena*. Furthermore, Cohn et al. (1984) studied an 831-bp region of the Adh gene from addi-

Table 1. *Drosophila melanogaster* Adh nucleotide sequence polymorphisms, and corresponding sequences of *D. simulans* and *D. mauritiana*

		Adh-s							Adh-f						
Region	Position	Consensus	1[a]	2	3	4	5	6	7	8	9	10	11	12	Partition
5′ Flanking	−3	C	.	.	.	.	.	.	.	T	T	T	T	.	a
	−2	C	.	.	.	.	.	.	.	G	G	G	G	.	a
	−1	G	.	C	.	.	.	C	C	C	C	C	C	.	b
Intron 1	107	C	.	.	.	.	A	.	.	A	A	A	A	.	c
	113	A	.	.	.	.	G	.	.	G	G	G	G	G	c (c′)
	143	A	.	.	.	.	.	.	.	.	.	.	G	.	d
	169	T	.	.	.	.	.	.	.	.	.	.	G	.	d
	173	A	.	.	.	.	.	.	.	.	.	.	G	.	d
	175	T	A	.	.	.	A	.	.	A	A	A	A	A	e
	287	G	T	.	.	.	.	.	.	.	.	.	.	.	f
	293	G	.	.	.	.	T	.	.	T	T	T	.	.	g
	304	G	.	.	.	.	C	.	.	C	C	C	.	.	g
	448	1[b]	−	−	−	−	−	−	−	+	+	+	+	−	a
	516	C	.	.	.	.	.	.	.	G	G	G	.	.	h
	551	2	−	−	−	−	−	−	−	+	+	+	−	−	h
	586	G	.	.	.	.	.	.	.	.	.	.	T	.	d
Larval leader	713	C	.	.	.	.	A	.	.	.	.	.	.	.	i
Exon 2	816	T	.	.	.	.	G	G	G	G	G	G	G	.	j
Intron 2	896	A	.	.	.	G	G	.	.	.	.	.	.	.	k
	925	C	.	.	.	T	T	.	.	.	.	.	.	.	k
Exon 3	1068	C	T	T	.	.	.	.	.	.	.	.	.	.	l
	1229	C	T	T	.	.	.	.	.	.	.	.	.	T	l (l′)
	1235	C	.	.	.	.	.	.	.	.	.	.	A	.	d
	1283	C	A	A	.	.	.	.	.	.	.	.	.	.	l
Intron 3	1354	G	C	C	.	.	.	.	.	.	.	.	.	C	l (l′)
	1362	G	A	A	.	.	.	.	.	.	.	.	.	.	l
	1388	A	.	.	.	.	.	.	.	.	.	.	G	.	d
	1400	A	T	T	.	.	.	.	.	.	.	.	.	T	l (l′)
	1405	T	A	A	.	.	.	.	.	.	.	.	.	.	l
Exon 4	1425	C	A	A	.	.	.	.	.	.	.	.	.	.	l
	1431	T	C	C	.	.	.	.	.	.	.	.	.	.	l
	1443	C	.	.	.	.	.	.	G	G	G	G	G	.	m
	1452	C	.	.	.	.	.	T	T	T	T	T	T	.	n
(Lys → Thr)	1490	A	.	.	.	.	.	.	C	C	C	C	C	.	m
	1518	C	.	.	.	.	.	T	T	T	T	T	T	.	n
	1527	T	.	.	.	.	.	.	C	C	C	C	C	C	m (m′)
	1557	A	.	.	.	.	.	C	C	C	C	C	C	C	n (n′)
	1596	G	.	.	A	A	.	A	.	.	.	.	.	.	o
3′ Untranslated	1693	A	.	.	.	.	C	C	C	C	C	C	C		j
	1698	3	1	1	1	0	4	5	5	5	5	6	5		i, j, p & q
	1740	C	.	.	.	.	.	.	.	G	G	G	.		h
3′ Flanking	1908	A	.	.	.	T	.	.	.	.	.	.	.		p
	1925	G	.	.	.	A	.	.	.	.	.	.	.		p
	1937	C	.	.	T	.	.	.	.	.	.	.	.		r
	2081	4	−	−	+	−	−	−	−	−	−	−	−		r
	2130	C	.	.	.	.	.	T	.	.	.	.	.		s
	2303	5	1	1	2	1	1	1	1	1	1	1	0		d & r
	2347	T	.	.	A	.	.	.	.	.	.	.	.		r
	2379	6	+	+	?	−	−	−	−	−	−	−	−		l

Data from Kreitman (1983), Bodmer and Ashburner (1984), and Cohn et al. (1984). Position 1490 is responsible for the fast–slow polymorphism. Exon 1 was not variable. Intron 1 is also called the adult intron. Af-s was not sequenced past position 2347

[a] Labels 1, 2, . . . , 11 correspond to Wa-s, Fl-1s, Af-s, Fr-s, Fl-2s, Ja-s, Fl-f, Fr-f, Wa-f, Af-f, and Ja-f, respectively, and 12 corresponds to the common sequence of *D. simulans* and *D. mauritiana*

[b] Labels 1, 2, . . . , 6 in the consensus Adh-s sequence refer to length polymorphisms. Presence (+) or absence (−) of an insertion is indicated, except for homonucleotide repeats (3 and 5), for which the excess above the minimum observed length is given

Table 2. Numbers of nucleotide differences (above diagonal) and estimates of the number of nucleotide differences per site (below diagonal) between 11 *Drosophila melanogaster* sequences

	Wa-s	Fl-1s	Af-s	Fr-s	Fl-2s	Ja-s	Fl-f	Fr-f	Wa-f	Af-f	Ja-f
Wa-s		3	14	16	19	19	20	27	27	27	29
Fl-1s	0.0013		13	15	20	16	17	26	26	26	28
Af-s	0.0059	0.0055		6	13	9	12	21	21	21	23
Fr-s	0.0068	0.0063	0.0025		11	11	14	23	23	23	25
Fl-2s	0.0080	0.0085	0.0055	0.0046		14	15	14	14	14	20
Ja-s	0.0080	0.0068	0.0038	0.0046	0.0059		5	14	14	14	16
Fl-f	0.0085	0.0072	0.0051	0.0059	0.0063	0.0021		9	9	9	11
Fr-f	0.0114	0.0110	0.0089	0.0097	0.0059	0.0059	0.0038		0	0	10
Wa-f	0.0114	0.0110	0.0089	0.0097	0.0059	0.0059	0.0038	0.000		0	10
Af-f	0.0114	0.0110	0.0089	0.0097	0.0059	0.0059	0.0038	0.000	0.000		10
Ja-f	0.0123	0.0119	0.0097	0.0106	0.0085	0.0068	0.0046	0.0042	0.0042	0.0042	

The number of nucleotide sites compared is 2379. Data from Kreitman (1983)

tional strains of *D. simulans* and *D. mauritiana*. The sequence from *D. orena* was not used in the present study because of difficulty in aligning this sequence with others. Figure 1 shows all differences among the three species used. We note that sites that were segregating in the *D. melanogaster* sample are all fixed in the four sequences sampled from *D. simulans* and *D. mauritiana*. Thus, variation within the *D. simulans* and *D. mauritiana* sample and variation differentiating this sample from *D. melanogaster* exists at sites distinct from those that vary within *D. melanogaster*. This implies that short-term evolution at the molecular level is not confined to a small number of sites, as has been implicated for satellite DNA (Brown and Clegg 1983) and mitochondrial DNA (Aquadro et al. 1984).

The study by Cohn et al. (1984) discloses 32 variable sites in the 831-bp region they sequenced. The study by Bodmer and Ashburner (1984) corroborates most of these variable sites, although the two studies differ in either sequence or alignment at several nucleotide sites. Naturally, most of the differences between the studies are rationalized as intraspecific polymorphisms, although at one site *D. simulans* and *D. mauritiana* apparently share the same polymorphism, and at two other sites *D. melanogaster* and *D. mauritiana* are fixed for different nucleotides, with *D. simulans* segregating for these nucleotides at both sites (arrows in Fig. 1). If *D. simulans* and *D. mauritiana* diverged relatively recently, it is reasonable that these two species share polymorphisms that existed prior to speciation. Several other differences between the two published versions have been corrected by corresponding with the respective authors, and are detailed in the legend to Fig. 1. Both *D. mauritiana* and *D. simulans* are apparently fixed for the amino acid corresponding to the *D. melanogaster* Adh-s electromorph. *D. mauritiana* differs from *D. melanogaster* Adh-s by five amino acid replacements, and from *D. simulans* by three replacements, whereas *D. simulans* differs by two amino acids from *D. melanogaster* Adh-s (Fig. 1).

Estimates of the number of nucleotide differences per site (d) within and between *D. melanogaster, D. simulans,* and *D. mauritiana* are shown in Table 3. The average number of nucleotide differences for all pairs of sequences within species is less than 1%. The number of net nucleotide differences per site between two species (Nei and Li 1979) was estimated by

$$\delta = d_{xy} - (d_x + d_y)/2 \qquad (3)$$

where d_x and d_y are the average numbers of nucleotide differences per site in each species, and d_{xy} is that between the two species. This formula corrects for the effect of the intraspecific polymorphism existing prior to speciation. Pairwise interspecific differences between *D. simulans* and *D. mauritiana* are less than those between these species and *D. melanogaster*. It should be noted that the number of net nucleotide differences ($\delta = 0.0086$) between *D. simulans* and *D. mauritiana* is about the same as the average uncorrected number of nucleotide differences ($d_{xy} = 0.0085$) between Adh-s and Adh-f (see also Zwiebel et al. 1982; Ashburner et al. 1984; Bodmer and Ashburner 1984). *D. simulans* and *D. mauritiana* sequences differ from *D. melanogaster* Adh-s sequences by an average of 0.0264 (uncorrected) substitutions per site and from *D. melanogaster* Adh-f sequences by 0.0278.

Phylogenetic Analyses

UPGMA Trees

We first constructed a phylogenetic tree for *D. melanogaster* sequences from the d values in Table 2 by using UPGMA. The tree obtained is presented

Table 3. Estimates of the number of nucleotide differences per site within and between *Drosophila melanogaster, D. simulans,* and *D. mauritiana*

	Mean ± SE[a]	Source[b]	Number of nucleotides
Nucleotide differences/site (d_x) within:			
D. mel. (total)	0.0066 ± 0.0017	K	2379
Adh-s	0.0056 ± 0.0015	K	2379
Adh-f	0.0028 ± 0.0011	K	2379
Adh-s vs. Adh-f	0.0085 ± 0.0019	K	2379
D. mel.	0.0070 ± 0.0041	K	822
D. sim.	0.0073 ± 0.0030	CTM, BA	822
D. mau.	0.0049 ± 0.0024	CTM, BA	822
UPGMA estimates of nucleotide differences between *D. mel.* and (*D. sim.* and *D. mau.*):			
Adh-s	0.0264 ± 0.0048	K, CTM, BA	822
Adh-f	0.0278 ± 0.0053	K, CTM, BA	822
Total	0.0270 ± 0.0049	K, CTM, BA	822
Interspecific comparisons (d_{xy}):			
D. mel. vs. *D. sim.*	0.0245 ± 0.0049	K, CTM, BA	822
D. mel. vs. *D. mau.*	0.0296 ± 0.0054	K, CTM, BA	822
D. sim. vs. *D. mau.*	0.0147 ± 0.0038	CTM, BA	822
Estimates of net nucleotide difference (δ):			
D. mel. vs. *D. sim.*	0.0173 ± 0.0028	K, CTM, BA	822
D. mel. vs. *D. mau.*	0.0236 ± 0.0024	K, CTM, BA	822
D. sim. vs. *D. mau.*	0.0086 ± 0.0022	CTM, BA	822
D. mel. vs. (*D. sim.* and *D. mau.*)	0.0205	K, CTM, BA	822

[a] The mean d_x was calculated by averaging values obtained by Eq. (1). The standard error of this mean was calculated by using Eq. (2) and assuming that there were only two sequences. This value is expected to overestimate the true variance when d_x is based on more than two sequences. UPGMA standard errors were calculated by the method of Nei et al. (1985). Standard errors of d_{xy} and δ were estimated by Eqs. (26) and (25) of Nei and Tajima (1981)

[b] K, Kreitman (1983); CTM, Cohn et al. (1984); BA, Bodmer and Ashburner (1984), restricted to the 822-bp region comparable to the CTM study

in Fig. 2. Sequences from France, Washington, and Africa (all Adh-f) were identical at the 43 polymorphic nucleotide sites, although an additional base in a homonucleotide run (site 1698) distinguishes Af-f from the other two. Nonetheless, their common sequence was used as three distinct operational taxonomic units.

The standard errors of branching points shown in Figs. 2 and 3 were obtained by the method of Nei et al. (1985). Factors such as intragenic recombination and gene conversion are not considered in the model underlying the standard error calculations. Since UPGMA is a method of phenetic clustering, i.e., clustering based on overall similarity, nucleotide differences caused by recombination will inflate the estimated standard errors of lower clusters. Conversely, standard errors estimated for upper branching points are generally underestimated because of similarities across clusters due to recombination.

The tree in Fig. 2 shows a basic division of the sequences corresponding to the fast–slow dichotomy. The branching point between Adh-s and Adh-f sequences is significantly older (the standardized normal deviate t = 2.42) than the oldest branching point within the Adh-f cluster, even though the Fl-f sequence is placed deep within the Adh-s cluster. It is also apparent that there are at least two major groups among the Adh-s sequences, one (S0) being represented by the Wa-s and Fl-1s sequences, and the other (S1) by the remaining Adh-s sequences plus Fl-f. Our statistical test (Nei et al. 1985) shows that the nucleotide sequence difference between the Ja-s and Fl-f sequences (d = 0.0021) is significantly smaller (t = 3.2) than the average difference (d = 0.0068) between S0 and S1. These significance calculations are not corrected for the inferred recombination events described later.

It should be noted that the UPGMA tree contains several ambiguous branching points. One of these is the joining of the two major Adh-s lineages (S0 and S1) prior to the final addition of the Adh-f group to the tree. This is due to the fact that the difference (d = 0.007138) between S0 and S1 is trivially smaller than that between S1 and the "fast" group (d = 0.007141). This ambiguity is adequately reflected by the standard errors of the upper branching points, which show graphically that we have no statistical confidence in the upper branching points among Adh-s sequences depicted in Fig. 2. Another potentially misleading cluster is that of Fl-f with Ja-s. The difference between these two sequences is only

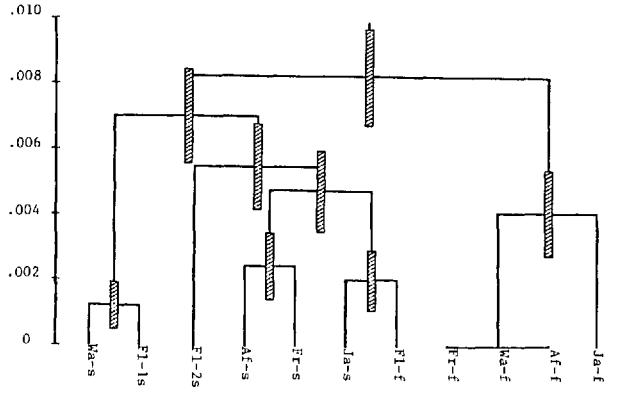

Fig. 2. UPGMA tree of 11 *D. melanogaster* sequences based on 43 segregating nucleotide sites in 2379 sequenced bases. The hatched bar for each branching point indicates one standard error on each side of the branching point. Scale is in number of nucleotide differences per site

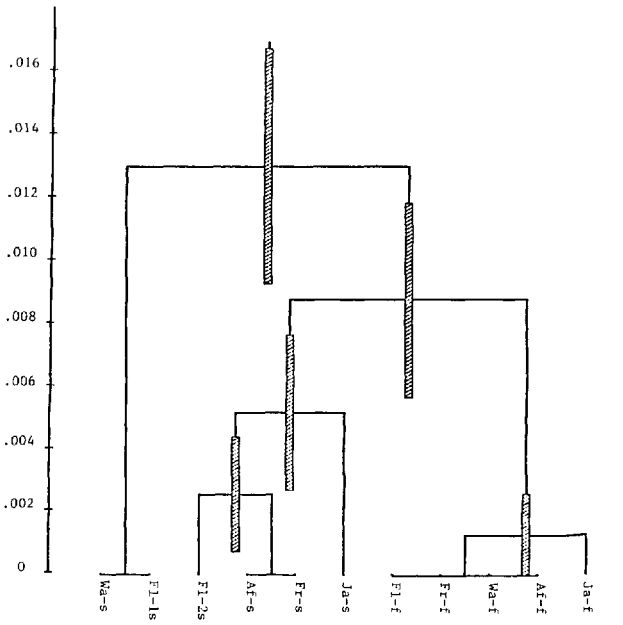

Fig. 3. UPGMA tree of 11 *D. melanogaster* sequences based on 14 segregating nucleotide sites in the 765 sequenced bases of the Adh coding region. Standard errors and scale as in Fig. 2

slightly smaller than the differences between Fl-f and the other Adh-f sequences (Table 2). Indeed, if we consider the coding region (765 bp) only, UPGMA gives an alternative tree in which all Adh-f sequences make one cluster and S0 clusters well outside (Fig. 3). Especially noteworthy is that the coding region of the Fl-f sequence is identical with that of the majority of Adh-f sequences.

The phylogenetic relationship of all 15 sequences (two each from *D. simulans* and *D. mauritiana*, plus the 11 *D. melanogaster* sequences) based on the 822 bp common to all shows that the divergence between *D. melanogaster* sequences and the *D. simulans–D. mauritiana* sequences is significantly earlier than the divergence between *D. simulans* and *D. mauritiana* sequences ($t = 4.77$). It should be noted that a tree based on amino acid differences would be misleading, because the number of amino acid substitutions is very small.

Parsimony Trees

Because of the ambiguity of the UPGMA tree, it was desirable to use parsimony methods (Fitch 1977) to obtain additional phylogenetic inferences. An important conceptual distinction should be mentioned at this point: For estimating expected branch lengths of a tree under the assumption of a molecular clock, UPGMA is the method of choice (Chakraborty 1977; Tateno et al. 1982; Nei et al. 1983). However, if the number of parallel and backward mutations is expected to be small and one wants to identify particular substitutions and in which branches they actually occurred, parsimony methods are expected to be better than UPGMA.

Figure 4 shows a parsimony tree obtained by Fitch's (1977) algorithm. This tree requires 63 mutational changes (53 point mutations and 10 length mutations) within the *D. melanogaster* sequences. This number is larger than the observed number of mutational changes by 10, but smaller than the number required for the UPGMA tree by 9.

For purposes of parsimony analysis and statistical testing of nucleotide site associations, all 49 polymorphic sites have been labeled (Table 1). The different labels correspond to specific partitionings of the sequences into two groups. For instance, three sites labeled m distinguish Adh-f from Adh-s. Each different partitioning is termed a phylogenetic partition (Stephens 1985). There are 1023 possible ways of partitioning the 11 *D. melanogaster* sequences, yet only 19 are realized. This is a strong indication that effects that would hinder phylogenetic analysis (e.g., parallel and backward mutations and recombination) are not prevalent. Furthermore, it is important to note that if 49 variable sites create only 19 different partitions, then most of the phylogenetic information (distribution among lineages) is redundant.

In the case of intraspecific sequence comparisons, the occurrence of the same mutational change in distantly related lineages may be caused by intragenic recombination or gene conversion. Hence, while parsimony methods minimize the total number of mutations, it may also be reasonable to ascribe some of the parallel changes in nucleotide sequences to these factors. We note that of the ten additional mutations required for the tree in Fig. 4, five are localized in intron 1 (c, e, and g in Table

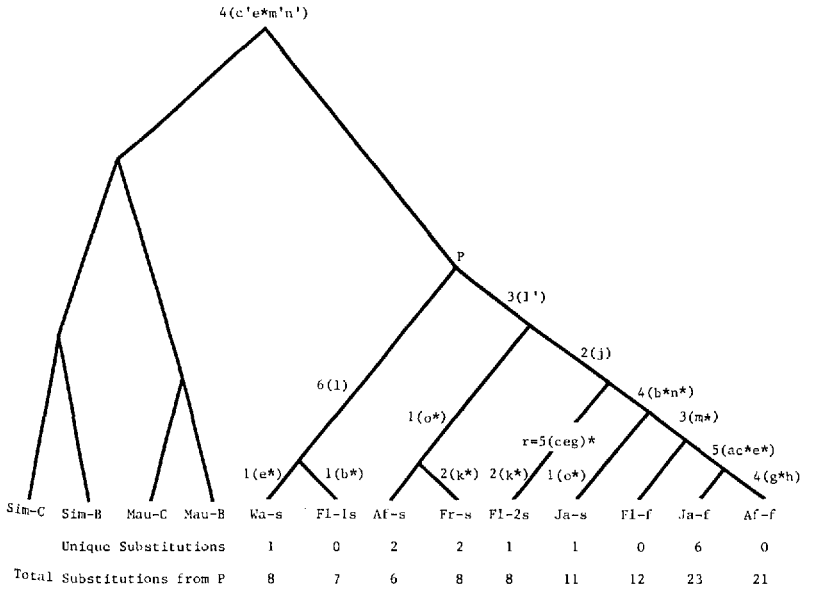

Fig. 4. Parsimony tree based on 30 phylogenetically informative (Fitch 1977) segregating nucleotide sites among the 43 determined by Kreitman (1983). Inferred phylogenetic changes at these sites (labeled a–s as in Table 1) are shown along branches. Numbers along branches indicate the number of sites involved; asterisks indicate multiple phylogenetic changes at a site (i.e., through parallel or backward mutations, or through recombination or gene conversion). Primes (') on c, l, m, and n indicate additional partitions created by including *D. simulans* and *D. mauritiana* sequences (see text). Intragenic recombination or gene conversion in the Fl-2s sequence, r = 5(ceg)*, replaces five mutations of conventional parsimony tree. See text for details

1). If we assume that there was no recombination and no gene conversion and that Fl-2s is legitimately within the Adh-s lineage, as in Figs. 2 and 4, the Fl-2s sequence of intron 1 can be explained only by five independent mutations. However, if we recognize this sequence as a recombinant, say, of sequences resembling Fr-s and Fr-f, then these five "mutations" can be considered to have occurred by a single recombinational event. Thus, only five additional mutations and one recombinational event need to be invoked for the phylogenetic tree of all *D. melanogaster* sequences in Fig. 4. Four additional nucleotide changes (c', e, m', n') are required for placing the *D. simulans* and *D. mauritiana* sequences on this tree, whether or not we regard Fl-2s as a recombinant.

The above considerations suggest that if the tree in Fig. 4 is correct, the Fl-2s sequence was produced either by intragenic recombination between a typical Adh-s sequence and a typical Adh-f sequence or by gene conversion involving a region (198–447 bp) of intron 1. Of course, it may not be that Fl-2s was "converted," but instead that the ancestral sequences of Ja-f, Af-f, Wa-f, and Fr-f were converted by Fl-2s by two separate events (first nucleotides labeled c and e, and later those labeled g). This possibility requires one additional recombinational event compared with the first interpretation, but the total number of mutational events required is still smaller than that for the case of no recombination or gene conversion.

As seen from Table 1, all length variations are well in accord with Fig. 4 and require no extra mutational events. Especially noteworthy is the indication that the length of homonucleotide repeats may change in a stepwise fashion.

A statistical method of analyzing the associations among nucleotide sites contributing to each partition (Stephens 1985) identifies at least six possible recombinational events. In Fig. 5, we have depicted the 11 sequences as combinations of segments of DNA that seem to have different evolutionary histories. Perhaps the most plausible recombination involves the Fl-1s sequence, which could, for instance, be a recombinant between sequences identical to Wa-s and Fl-f. Although Kreitman (1983) suggested the possible involvement of the Fl-1s sequence in a recombinational event, this would not be deduced by standard UPGMA or parsimony treatment of the data. Both of these methods treat the three nucleotide differences between Wa-s and Fl-1s as three independent evolutionary events.

Mutations specific to Af-s (partition r) and Fr-s (partition p) are significantly clustered at the 3′ ends of these sequences. In both cases, it is possible that the sequence is a recombinant with some unsampled relict sequence. The two remaining differences between Af-s and Fr-s both correspond to partition k. These two sites show that the sequences of intron 2 of Fr-s and Fl-2s are identical to each other, but different from all other sequences. We infer that a recent ancestor of Fr-s was converted by a sequence similar to Fl-2s, since Fl-2s has additional unique variation nearby (partition i). The sequence of Fl-2s may also have been converted by an Adh-f sequence, as was described above (Fig. 4).

The five differences between Ja-s and Fl-f are all in the 3′ half of the sequence. The probability that five differences would span a distance as short or shorter than that observed is only 0.0162. We infer that, since many of the sites 5′ of this region are variable (Fig. 1), the clustering is due to one of the

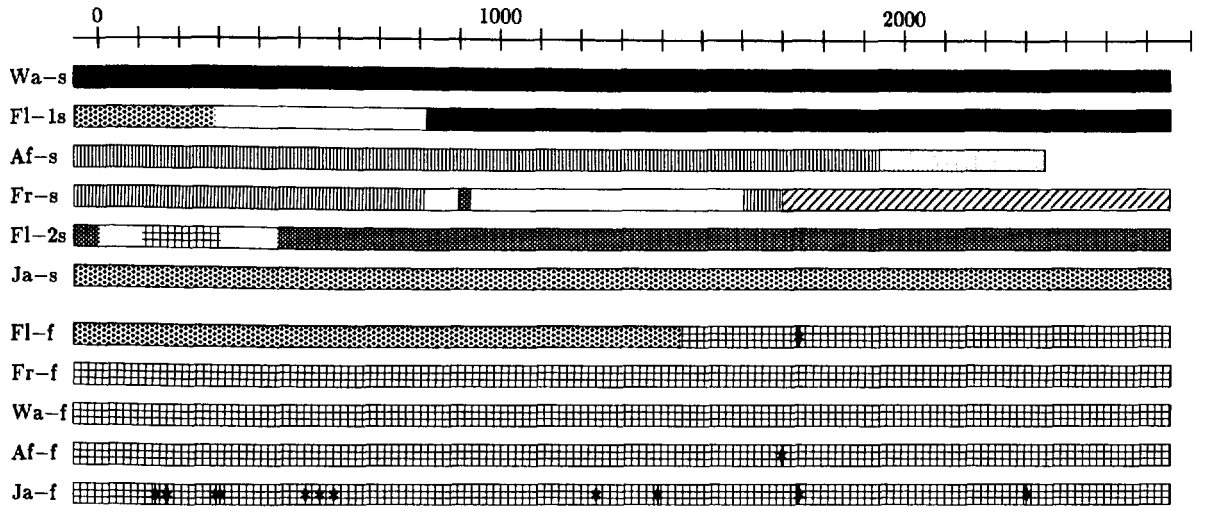

Fig. 5. Representation of *D. melanogaster* sequences as combinations of DNA segments with different evolutionary histories. Similar shading indicates sequence identity, with differences indicated by asterisks. Open areas indicate lack of diagnostic sites for determining end points of recombination events. Scale is that used in Table 1 and Kreitman (1983)

sequences being a recombinant, probably Fl-f (Fig. 5).

The above account of the evolutionary history of the Adh sequences calls for a total of 64 events (53 observed mutations, 6 recombinational events, and 5 parallel or backward mutations, including 3 inferred parallel changes in the sibling species). The "phylogenetic tree" of the different sequences and segments is presented in Fig. 6. Note that we treat the 3′ ends of the Af-s and Fr-s sequences as distinct, ancient lineages, since they are highly diverged in an otherwise highly conserved region. This tree represents our inferences about the evolutionary history of the Adh sequences and sequence segments when we are allowed to invoke recombination and gene conversion with the observed statistical associations of nucleotide polymorphisms found by applying Stephens's (1985) method. Statistical association of nucleotide polymorphisms may also occur when there are "hot spots" of recombination (see Kazazian et al. 1983) and biased nucleotide substitution such as a high frequency of transitional nucleotide substitution. Therefore, it should be kept in mind that the tree in Fig. 6 may not necessarily be better than that in Fig. 4.

Comparisons with Sibling Species

When information from *D. simulans* and *D. mauritiana* (Bodmer and Ashburner 1984; Cohn et al. 1984) is included, four new partitions are created from partitions c, l, m, and n. For instance, the four sequences from *D. simulans* and *D. mauritiana* have the same nucleotide as the Adh-s sequences at two sites (partition m), but they have the same nucleotide as the Adh-f sequences at a third site (1527), which creates the partition labeled m′. Similarly, partitions c′, l′, and n′ reflect that the relevant nucleotides are not those of the consensus Adh-s sequence.

Bodmer and Ashburner's (1984) sequences support the view that Kreitman's (1983) consensus Adh-s sequence is the ancestral sequence at 29 of the 36 variable sites that are overlapped (Table 1). Four (e and l′, Table 1) of the other seven may be ancestral (Fig. 6), although the other three (c′, m′, and n′) require additional mutations. All seven polymorphic sites at which *D. simulans* and *D. mauritiana* differ from the *D. melanogaster* consensus "slow" sequence carry the variant observed by Kreitman. Kreitman's study demonstrated seven more polymorphic sites in an additional 978 bp immediately downstream of the coding region, but most of these polymorphisms were unique, suggesting that the consensus slow sequence is also the common ancestor of this region.

Times of Divergence of DNA Sequences and Species

The rate of nucleotide substitution in evolution seems to be approximately constant (Kimura 1983; Li et al. 1985). Therefore, we can estimate the time of divergence between DNA sequences using the equation $d = 2\lambda t$, where λ and t are the rate of nucleotide substitution per site per year in one lineage and the time measured in years, respectively. We shall use two values of λ, $\lambda_1 = 3.0 \times 10^{-9}$ and $\lambda_2 = 5.0 \times 10^{-9}$, which are estimates of the rates for introns and silent sites, respectively (Li and Gojobori

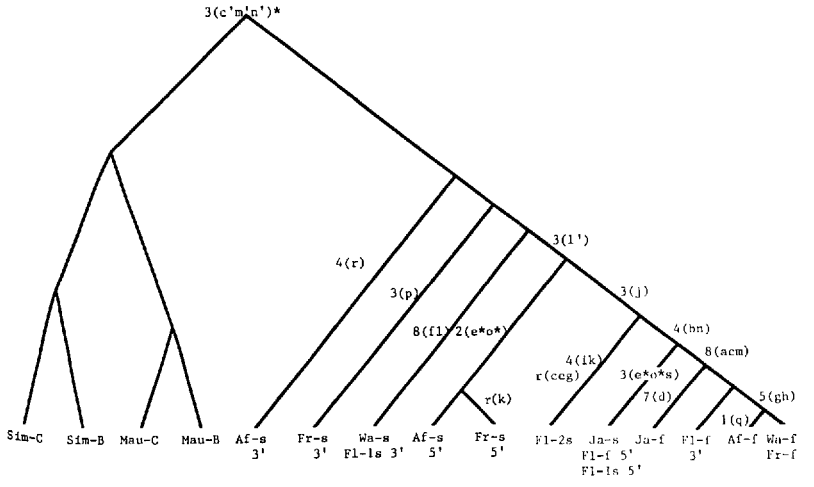

Fig. 6. An evolutionary history of the sequences and segments in Fig. 5. Each labeled end point presumably reflects a different lineage, and it is the evolutionary relationship of these lineages that is being described. Nucleotide changes and embedded recombination events are shown as in Fig. 4, except asterisks no longer refer to recombination

1984). These estimates are representative of the range of many estimates of substitution rates (Kimura 1983; Li et al. 1985), including one for *Drosophila* ($\lambda = 3.3 \times 10^{-9}$) based on DNA–DNA hybridization (Zwiebel et al. 1982). We apply λ_1 and λ_2 to all sites, and treat the small fraction of nonsynonymous substitutions observed as if they were silent. Kreitman (1983) identified 573 bp of the coding region as "non-silent," which means that if we include these sites, our estimates of divergence time will be underestimates, since nonsilent sites are known to be generally quite conservative.

We first estimate the coalescence time of all sequences sampled from *D. melanogaster*, i.e., the time at which all sampled sequences trace back to a common ancestral sequence. If we use the UPGMA tree in Fig. 2, this time is 820,000 ± 145,000 years when λ_2 is used, and 1,360,000 ± 240,000 years when λ_1 is used. However, if we use the tree in Fig. 3, it becomes 1.3–2.0 million years (Myr). This indicates that the polymorphic sequences at the Adh locus are very old. Antiquity of polymorphic alleles is expected under the neutral mutation theory, as will be discussed later. In fact, the above calculations may be underestimates. While reciprocal intragenic recombination has no effect on the expected number of nucleotide differences between two randomly chosen sequences, the variance is expected to decrease (Kimura 1969; Watterson 1975; Hudson 1983). Hence the maximum difference among all pairs of sequences may be smaller than in the case of no recombination.

The age of the fast–slow electrophoretic polymorphism is of special interest, because this polymorphism has been studied extensively. Obviously, this polymorphism must have arisen between the splitting time between the Adh-f and Adh-s sequences and the splitting time of the two most divergent Adh-f sequences. We will use the splitting time between the Adh-f and Adh-s sequences as an upper bound for the age of the electrophoretic polymorphism. If we use the phylogenetic tree in Fig. 2, which treats Fl-f as an Adh-s sequence, the time of the fast–slow split is identical with the coalescence time for all sequences, which we estimated above as 820,000–1,360,000 years. This estimate changes only slightly if we use the phylogenetic tree in Fig. 3: In this case the estimate of the time of the fast–slow split is 880,000–1,470,000 years. In the above two computations, nonsilent sites in the coding region were treated as though they were silent. If we exclude these nonsilent sites from the computations, our estimates of the fast–slow splitting time become 30% greater for the tree in Fig. 2 and quadruple for the tree in Fig. 3. (In this case we use λ_2 only.) Kreitman (1983) noted a statistical difference in the level of silent polymorphism: That of the coding region was higher than that in the introns. Easteal and Oakeshott (1985) have recently reanalyzed Bodmer and Ashburner's data (including those for *D. orena*) and concluded that the highly conserved regions, rather than the rapidly evolving regions, are atypical, which would seem to indicate that the more ancient splitting times obtained by excluding nonsilent sites are more reasonable.

Another estimate of the age of the fast–slow polymorphism is obtained by using Ja-s, which is the Adh-s sequence most closely related to the Adh-f sequences (Fig. 4). The average d value between this sequence and Adh-f sequences is 0.0061, if we exclude Fl-f because of the possibility of recombination. Therefore, the time of the fast–slow split is estimated to be 610,000–1,000,000 years ago. If we exclude nonsilent sites in this case, the estimates again increase by 30%. Our estimates of the age of the fast–slow polymorphism cover a wide range, from 610,000 years to 3.5 Myr. We feel that 1 Myr is perhaps the best rough estimate available. In any

case the polymorphism would seem to be at least a few hundred thousand years old, and potentially much older.

The intraspecific nucleotide differences in *D. simulans* and *D. mauritiana* are 0.0073 and 0.0049, respectively (Table 3), corresponding to divergence times of 730,000–1,200,000 and 490,000–800,000 years, respectively. Correction for nonsilent sites and substitutions almost triples the estimate for *D. simulans*, but only slightly increases that for *D. mauritiana*.

Cohn et al. (1984), Bodmer and Ashburner (1984), and Ashburner et al. (1984) estimated the times of species divergence for pairs of *D. melanogaster, D. simulans*, and *D. mauritiana*. Their estimates, based on λ values essentially the same as ours, range from 2.7 to 4.7 Myr. However, they did not correct for polymorphism within species, as each of their studies was based on single sequences from each species. To correct for this effect, we use the number (δ) of net nucleotide differences. If we use the δ values given in Table 3, the estimates of the time of divergence between *D. simulans* and *D. mauritiana* become 0.86–1.45 Myr ago, whereas the estimate for divergence between *D. melanogaster* and these two species is 2.0–3.5 Myr ago. These estimates are considerably smaller than the estimates obtained by the previous authors. Furthermore, the study by Easteal and Oakeshott (1985) excluded three apparently conserved regions, which inflated their estimates to 7.7 and 9.4 Myr, respectively. In the present case it is clearly important to consider the effect of polymorphism.

Discussion

Kreitman's finding that Adh-f sequences from diverse geographic localities are more similar to each other than to Adh-s sequences from sympatric flies is consistent with the theory that Adh-f is a more recent allele, of single origin. That three variable nucleotide sites (m, Table 1) are diagnostic for the fast–slow difference would seem to establish a single origin for the worldwide majority of Adh-f sequences. Furthermore, Ashburner et al. (1984) have shown that most of the nucleotide substitutions common in the Adh-f sequences (including the amino acid variant) do not appear in the sibling species, and appear only sporadically in Adh-s, suggesting that Adh-f is recently derived.

We have seen that in the evolution of polymorphic sequences of the Adh locus intragenic recombination or gene conversion apparently played a significant role. Although this possibility was previously indicated by Kreitman, our phylogenetic analysis gives a clearer picture of the role of these evolutionary mechanisms. In the present paper we considered only those cases involving several polymorphic nucleotide sites. Clearly, if the breakpoints of recombination are close together, such events will be indistinguishable from parallel mutations.

In the presence of these mechanisms, the reconstruction of phylogenetic trees becomes quite complicated. If the frequency of occurrence of recombinations were extremely high, it would be almost impossible to reconstruct a phylogenetic tree of polymorphic sequences from the same species. In the present case, however, the frequency does not seem to be so high as to make phylogenetic reconstruction useless, since in our analysis the fast–slow allelic dichotomy has been clearly observed, and only two parallel mutations are needed for the *D. melanogaster* sequences in the tree of Fig. 6. When one is interested in constructing a phylogenetic tree of genes sampled from distantly related species, this problem is not serious unless the gene under investigation belongs to a multigene family.

We note that in the tree in Fig. 4 the total number of mutational events from the ancestor P varies considerably with DNA sequence, the average number for the Adh-f sequences being about 3 times higher than that for the Adh-s sequences. This seems to suggest that the Adh-f sequences have changed much faster than the Adh-s sequences. However, this is partly due to properties of the parsimony method used. In general, parsimony methods tend to assign more mutations to short internodes than to long "legs" of the tree. This is because two independent parallel mutations occurring in two different legs can always be explained by a single mutation occurring in an internode immediately ancestral to the two legs. The probability of having undetectable backward and parallel mutations is also higher in long legs than in short internodes or short legs. In Fig. 4, however, the major factor contributing to the difference between Adh-f and Adh-s alleles could be improper or incomplete assignment of intragenic recombination or gene conversion. As mentioned earlier, in the tree in Fig. 4, Fl-2s could be a donor rather than a recipient of the "converted" gene segment. If this type of gene conversion or intragenic recombination occurred several times in the past, the difference in the rate of accumulation of mutational changes between the Adh-f and Adh-s alleles would be reduced substantially.

Nevertheless, the Adh-f sequences seem to have accumulated mutations somewhat faster than the Adh-s sequences. This is seen by comparing the Adh-f sequences (A) and the Adh-s sequences (B) with the *D. simulans* and *D. mauritiana* sequences (C). The uncorrected average numbers of nucleotide differences among these three groups of sequences are $d_{AB} = 0.0085$, $d_{AC} = 0.0278$, and $d_{BC} = 0.0264$, where subscripts A, B, and C refer to the groups

concerned (Table 3). Following Fitch and Margoliash (1967), we can estimate the average distance (a) between P in Fig. 4 and the Adh-f sequences and the average distance (b) between P and the Adh-s sequences using d_{AB}, d_{AC}, and d_{BC}. The values are $a = 0.0050$ and $b = 0.0036$. Thus, a is about 50% larger than b. However, this difference is not statistically significant when a and b are assumed to be Poisson variables. Therefore, the difference could be due to stochastic errors.

The stochastic variance of the number of nucleotide differences is known to be very large. If we use the infinite-site model of neutral mutations, this variance, including the sampling variance, can be evaluated by using formulae developed by Tajima (1983) and Takahata and Nei (1985). For example, we previously estimated the average number of nucleotide differences per site (π) for the 11 *D. melanogaster* sequences to be 0.0066. If we use Tajima's formula (30), the expected standard error of this estimate under the infinite-site model becomes 0.0036, which is roughly half as large as the estimate itself. The estimated number of net nucleotide differences (δ) also usually has a large standard error when δ is small (Takahata and Nei 1985). For example, in the case of *D. simulans* and *D. mauritiana* we obtained $\delta = 0.0086$, but the expected standard error is 0.0075. This large standard error is due partially to the use of only two sequences from each species, but even if a large sample size is used, the standard error remains appreciably high. In the case of *D. melanogaster* vs *D. simulans* the mean and standard error become $\delta = 0.0173 \pm 0.0090$, whereas the corresponding values for the case of *D. melanogaster* vs *D. mauritiana* are 0.0236 ± 0.0086. In these cases the expected standard error is about one half the estimate. Note that in the above evaluation of the variance of π or δ the effects of recombination and gene conversion are ignored.

Kreitman's data present a unique opportunity to check the consistency of several aspects of the neutral theory. Under the neutral theory, the average nucleotide difference per site (π) in an equilibrium population should be equal to $4N_e\mu$, where μ is the mutation rate per site per generation and N_e is the effective population size. Thus, if we assume that there are six generations in a year in *Drosophila* and the mutation rate is equal to the substitution rate, we have $\pi = 4N_e\lambda_2/6 = 0.0066$ from Table 3. This gives $N_e = 2 \times 10^6$, a value roughly similar to Kreitman's estimate (3.3×10^6) based on silent sites of the coding region alone. The expected time of coalescence of r sequences (T_r), i.e., the expected time of divergence of r sequences from a common ancestor, is $4N_e(1 - 1/r)$ generations (Kingman 1982; Tajima 1983). If we use $N_e = 2 \times 10^6$ and $r = 11$, T_r becomes 7.3×10^6 generations. Again assuming six generations per year, the age of the oldest polymorphism in the sample is thus estimated to be 1.2×10^6 years old, which is very close to the estimate from the number of nucleotide differences (Table 2). Note that our estimate of the true age of the oldest polymorphism depends on our choice of λ, but the ratio of T_r to the oldest split time does not. Note also that the theory used above depends on the assumption of no recombination and no gene conversion. In the present case, however, the effects of these factors seem to be rather small, since the phylogenetic trees in Figs. 2, 4, and 6 are all in rough agreement with each other. Thus, the agreement between the two estimates of coalescence time indicates that the data on nucleotide polymorphism and evolutionary divergence in *D. melanogaster* are consistent with the predictions from the neutral theory. Of course, this set of data may be explained by some other hypotheses invoking selection as well.

Acknowledgments. We would like to thank C. Aquadro, T. Gojobori, M. Kreitman, W.-H. Li, N. Saitou, and N. Takahata for their comments. We also thank M. Ashburner and V. Cohn for access to their unpublished data and for their aid in reconciling their sequences. This work was supported by research grants NIH GM 20293 and NSF BSR 83115.

References

Aquadro CF, Kaplan N, Risko KJ (1984) An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. Mol Biol Evol 1:423–434

Ashburner M, Bodmer M, Lemeunier F (1984) On the evolutionary relationships of *Drosophila melanogaster*. Dev Genet 4:295–312

Bodmer M, Ashburner M (1984) Conservation and change in the DNA sequence coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature 309:425–430

Brown AHD, Clegg MT (1983) Analysis of variation in related DNA sequences. In: Weir BS (ed) Statistical analysis of DNA sequence data. Marcel Dekker, New York, pp 107–132

Chakraborty R (1977) Estimation of time of divergence from phylogenetic studies. Can J Genet Cytol 19:217–223

Cohn VH, Thompson MA, Moore GP (1984) Nucleotide sequence comparison of the Adh gene in three drosophilids. J Mol Evol 20:31–37

Easteal S, Oakeshott JG (1985) Estimating divergence times of *Drosophila* species from DNA sequence comparisons. Mol Biol Evol 2:87–91

Fitch WM (1977) On the problem of discovering the most parsimonious tree. Am Nat 111:223–257

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Johnson FM, Schaffer HE (1973) Isozyme variability in species of the genus *Drosophila*. VII. Genotype-environment relationships in populations of *Drosophila melanogaster* from the eastern United States. Biochem Genet 10:149–163

Jukes TH, Cantor CH (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kazazian HH Jr, Chakravarti A, Orkin SH, Antonarakis SE (1983) DNA polymorphisms in the human β globin gene cluster. In: Nei M, Koehn RK (ed) Evolution of genes and proteins. Sinauer Associates, Sunderland, Massachusetts, pp 137–146

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, London

Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2:87–90

Kingman JFC (1982) On the genealogy of large populations. J Appl Prob 19A:27–43

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

Li W-H, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol 1:94–108

Li W-H, Luo C-C, Wu C-I (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum Press, New York, pp 1–94

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. Genetics 97:145–163

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol 19:153–170

Nei M, Stephens JC, Saitou N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol Biol Evol 2:66–85

Oakeshott JG, Gibson JB, Anderson PR, Knibb WR, Anderson DG, Chambers GK (1982) Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on different continents. Evolution 36:86–96

Stephens JC (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:323–344

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol 18:387–404

Thatcher DR (1980) The complete amino acid sequence of three alcohol dehydrogenase alleloenzymes (Adh^{N-11}, Adh^{S} and Adh^{UF}) from the fruitfly *Drosophila melanogaster*. Biochem J 187:875–886

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7: 256–276

Zwiebel LJ, Cohn VH, Wright DR, Moore GP (1982) Evolution of single-copy DNA and the ADH gene in seven drosophilids. J Mol Evol 19:62–71

Received June 12, 1985/Revised August 5, 1985