Nucleotide Polymorphism at the Xanthine Dehydrogenase Locus in *Drosophila pseudoobscura*[1]

Margaret A. Riley,[*,2] *Susanne R. Kaplan,*[*,3] *and Michel Veuille*[†]

[*]Museum of Comparative Zoology, Harvard University; and [†]Laboratoire de Biologie et Génétique Evolutives, CNRS

Sequential polyacrylamide electrophoresis has revealed 20 allozymes of xanthine dehydrogenase (XDH) in *Drosophila pseudoobscura.* DNA sequence determination of seven isolates of the *Xdh* locus that represent six allozyme classes are presented here. Of the 5,456 sites examined, 180 are polymorphic, with 27 polymorphisms occurring at nonsynonymous, or replacement, sites. An average of nine amino acids differ between XDH allozyme classes, with 85% of the polymorphic amino acids singly represented. The level and pattern of variation observed at *Xdh* argue that the effective population size of the species is quite large—i.e., on the order of 2×10^6—and that the populations sampled are quite ancient. In addition, as judged by two statistical tests, the levels of nucleotide polymorphism observed at *Xdh* are compatible with predictions from the neutral theory of molecular evolution.

Introduction

It has been argued that DNA sequence surveys will provide the sort of data required to test hypotheses regarding the forces operating on genetic variation (Lewontin 1985; Kreitman 1988). The power of this approach lies both in the complete resolution, afforded by DNA sequence information, of molecular polymorphism and in the contrast between levels of polymorphism and divergence at closely linked but functionally distinct nucleotide sites, e.g., synonymous versus nonsynonymous positions. It is possible, by this approach, to distinguish between those loci that are evolving in a neutral fashion, i.e., under the primary influence of neutral mutation and random genetic drift, and those clearly under the influence of some form of positive selection (for a review of neutral theory, see Kimura 1983, pp. 34–55; for specific tests of neutrality, see Hudson et al. 1987; Tajima 1989).

A large survey of DNA sequence variation at the alcohol dehydrogenase (ADH) locus (*Adh*) in *Drosophila melanogaster* has clearly revealed the power of this approach. A combination of complete DNA sequences for 11 *Adh* genes (Kreitman 1983), one sequence from a closely related species, *D. simulans* (Kreitman and Aquade 1986*a*), and four-cutter restriction mapping for several hundred additional *Adh* genes (Kreitman and Aguade 1986*a,* 1986*b*; Simmons et al. 1989) have provided strong evidence for the role of natural selection in maintaining a single segregating amino acid as a balanced polymorphism at ADH (Hudson et al. 1987).

Adh was initially chosen for study by Kreitman (1980) because the locus dis-

[1] Key words: *Xdh, Drosophila,* nucleotide polymorphism, amino acid polymorphism.

[2] Present address: Department of Biology, Yale University, New Haven, Connecticut 06511.

[3] Present address: Department of Ecology and Evolution, University of Arizona, Tucson, Arizona 86721.

Address for correspondence and reprints: Margaret Riley, Department of Biology, Yale University, New Haven, Connecticut 06511.

Mol. Biol. Evol. 9(1):56–69. 1992.

0737-4038/92/0901-0005$02.00

played a widespread two-allele amino acid polymorphism in *D. melanogaster* populations. A repeated clinal distribution for these two major variants of ADH argued a priori for the importance of positive selection in producing the observed allozyme variation. The question remains, however, whether ADH is truly representative of the majority of protein polymorphism segregating in natural populations. Proponents of the neutral theory maintain that most electrophoretically distinguished protein variation represents selectively neutral alternatives in various stages of fixation or loss due to stochastic processes (Kimura 1983, p. 253).

We present here a sample of nucleotide variation for the xanthine dehydrogenase (XDH) locus (*Xdh*) in *D. pseudoobscura*. The present investigation is part of a larger study of sequence variation at the locus designed to discriminate among the forces controlling genetic variation at *Xdh* (Riley 1989; Riley et al. 1989). In contrast to ADH, the encoded enzyme XDH is highly polymorphic in all species of *Drosophila* where it has been examined electrophoretically (Coyne 1976; Buchanon and Johnson 1983; Keith et al. 1985). Keith et al. (1985), in a survey of XDH electrophoretic variation, revealed 20 allozymes in a sample of 184 genes from two populations of *D. pseudoobscura*. In addition, the two populations, while separated by >500 km, were shown to be statistically indistinguishable in allozyme identity and frequency distribution.

Several questions have been raised by electrophoretic surveys of variation at XDH: (1) What is the true nature and extent of amino acid variation revealed electrophoretically? Has sequential protein electrophoresis provided an accurate measure of this amino acid variation? (2) What evolutionary forces are responsible for the high level of amino acid variation observed? Are the 20 allozyme classes revealed by Keith et al. (1985) maintained in the populations by some form of positive selection, or are they selectively neutral alternatives segregating under the primary influence of random genetic drift?

In the present report we describe seven complete DNA sequences for the *Xdh* locus that represent six of the allozyme classes distinguished by Keith et al. (1985). These data reveal an unexpectedly high level of inferred amino acid polymorphism. Both the level and the pattern of this polymorphism would not have been predicted on the basis of previous electrophoretic studies. In addition, by contrasting the levels of polymorphism and divergence at synonymous and nonsynonymous sites, these data provide some insight into the importance of neutral versus selective forces operating at *Xdh*.

Material and Methods

The genomes for the present study were collected by Keith (1983) in May 1979, at the James Reserve (JR) in the San Jacinto Mountains in southern California and near the Gundlach-Bundschu winery (GB) in the Sonoma Valley of northern California. These locations are >500 km apart and differ in elevation by >1,500 m (JR 1,646 m; GB 30 m). Second chromosomes were extracted, and isochromosomal lines were maintained according to methods described by Keith et al. (1985). Approximately 2.5 g of each of seven fly lines chosen for the present study and described in table 1 were frozen in liquid nitrogen and stored at −70°C (in June 1986). Table 1 includes a description of each line in terms of both population origin and electrophoretic mobility [as determined by Keith et al. (1985)]. These lines were chosen to include a range of electrophoretic mobility variants—and thus they do not represent a random sample of genomes from the Keith et al. survey.

The *Xdh* region of fly line JR436 was cloned and sequenced according to methods

Table 1
Sequential Electrophoretic Mobility Classes of XDH for *Drosophila pseudoobscura* Isolates Included in Present Study

Sequential Electrophoresis Buffer Conditions[a]	Population Frequency[b] JR	Population Frequency[b] GB	Fly Line Designation
0.92/0.98/0.98/1.00/1.00	1	0	JR48
0.94/1.00/1.00/1.00/1.00	4	2	GB336
0.94/1.03/1.00/1.02/1.01	0	1	GB50
1.00/1.00/1.00/1.00/1.00	58	52	GB87+GB369
1.00/1.02/1.01/1.02/1.03	9	4	JR436
1.04/1.00/1.00/1.00/1.00	0	2	GB361

[a] Notation is that of Keith et al. (1985). XDH mobility under each of five buffer conditions is separated by a slash (/).

[b] Source: Keith et al. (1985).

described by Riley (1989). The six remaining fly lines were cloned and sequenced according to the following procedures: The *Xdh* region was localized, from genomic DNA, to an ~15-kb *Hin*dIII fragment by following standard Southern analysis procedures (Maniatis et al. 1982) and employing as a probe a 9-kb portion of the *Xdh* region from JR436, described by Riley (1989). A modification of standard lambda cloning procedures (Maniatis et al. 1982) was used to clone the 15-kb *Xdh* region from each fly line into *Hin*dIII-cut plasmid, pEMBL19 (Dente et al. 1983). The primary modification involved the utilization of a 5%–30% sucrose gradient to enrich for 15-kb-sized fragments (D. Curtis, personal communication). In addition, KH802 cells were employed in plasmid transformation, and the Hanahan (1983) procedure was utilized in making the cells competent. Approximately 200,000 recombinant colonies per plasmid library were screened (Maniatis et al. 1982) by employing the same 9-kb *Xdh* probe described above.

The *Xdh* locus and 196-bp 5′ noncoding region (see fig. 1) were sequenced in each of the additional six fly lines. Forty-seven oligonucleotide sequences (primers) 15–20 bp in length were synthesized. The primers cover both strands of a 5,456-bp region (bp 890–6345 in Riley 1989). Double-stranded plasmid sequencing was employed. Plasmids were transformed into DH5α cells, and miniprep plasmid DNA was prepared according to the alkaline procedure (Maniatis et al. 1982). Two micrograms of DNA and 10 ng of each primer were sequenced using the dideoxy chain-termination method of Sanger et al. (1977) according to the Sequenase protocol (Tabor and Richardson 1987).

Results

Nucleotide Polymorphism

The *Xdh* region, 5,456 bp in total, was sequenced in each of seven natural isolates of *Drosophila pseudoobscura*. Figure 2 lists the 185 nucleotide polymorphisms and six insertion/deletion variants. The base-pair numbering corresponds to that of the original *Xdh* sequence presented by Riley (1989). Table 2 provides two descriptions of the variation observed at *Xdh:* (1) nucleotide polymorphism and (2) nucleotide diversity, or base-pair heterozygosity (Nei 1987, p. 256).

Estimates of polymorphism are highly dependent on sample size. Indeed, in an

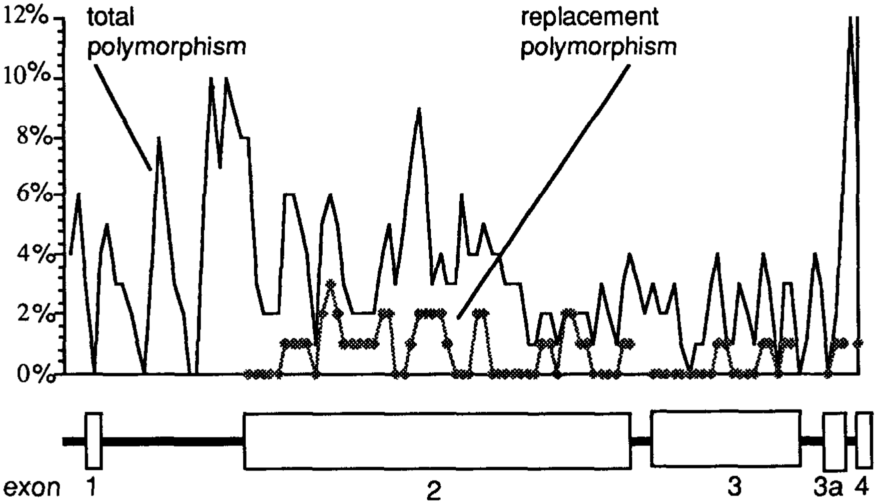

FIG. 1.—Percent nucleotide polymorphism at *Xdh* locus in *Drosophila pseudoobscura.* The physical map of the region is indicated along the *x*-axis, and % polymorphism, per 100 nucleotides, is given along the *y*-axis (a sliding window of 50 nucleotides was employed). Total polymorphism (intron, silent, and replacement) and replacement-site polymorphism are given separately.

infinite population all sites examined would be polymorphic. Estimates of nucleotide diversity are not as sensitive to sample size. However, of obvious interest in an analysis of sequence variation is a comparison, within each functional class of nucleotide sites, of the levels of variation distributed along the region sequenced—e.g., a comparison of either (1) levels of diversity at synonymous sites between the exons of *Xdh* or (2) levels of nonsynonymous-to-synonymous variation along the coding region. For fixed sample sizes, polymorphism is clearly correlated with nucleotide diversity. In addition, results from a high-resolution mapping survey of *Xdh* nucleotide variation argue that the polymorphic sites are in linkage equilibrium (Riley et al. 1989); thus the sites can be assumed to be evolving independently, and the variance in the number of segregating sites will be Poisson distributed. With these assumptions, χ^2 tests comparing levels of polymorphism can be employed.

Distribution of Polymorphism across the Coding Region and between Functional Classes of Sites

The nucleotide variation observed at *Xdh* is not randomly distributed. A goodness-of-fit test comparing the total level of polymorphism along the region sequenced reveals significant heterogeneity in distribution of variation across exons, introns, and 5′ and 3′ noncoding sequence ($G = 21.79$, degrees of freedom (df) = 5, $P \ll 0.01$, in an analysis that considers six equal blocks of sequence). Tests of heterogeneity across the coding sequence alone, by dividing the coding region either into six equal sized blocks or into exons, reveals that synonymous and replacement polymorphism, in total, is significantly clustered in distribution ($G = 526.9$, df = 4, $P \ll 0.001$, by exons; $G = 14.59$, df = 5, $P < 0.02$, by blocks), with an apparent excess of polymorphic sites in the 5′ end of exon 2. Nonsynonymous polymorphisms, when examined independently, are not significantly clustered ($G = 6.5$, df = 4, $P > 0.1$, by exons; $G = 7.94$, df = 5, $P > 0.1$, by blocks), although there are two times as many replacements in the 5′ end of exon 2 compared with any other block of coding sequence (see fig. 1).

```
                  111111111111111 1 1 11111 1111111111111222  22222222
               899900011112223 3555 5 5 55566 6888888999999000  00000011
               9566004557805725013 3 6 77735 6678899113488 9113  34447800
               9616014457205317220 3 6 02331 2014558017078 9385  91591201
Consensus      CCTACACGGGCGTGTTAT      TCGA  GAGGTACCCTTGAGGC  TGCTACAG
JR436          T.....T.........GA.............G..C......................
GB050          ...........A.A...............TG..C................CGT...
JR048          .......AAAA...........d8....i8.G......GA....A.....AA.....
GB336          ...CGG...AA...CAGA......CAC.....A.......C....A.........G.
GB361          T..........AA....A.......................TTA..i10C......A
GB087          .TA........A......i8d2.....T.....C..T.GA....A.....AA.....
GB369          .TA...............i8d2...........C.TTGGA....A.....AA.....
               [--1--][-------------------2----------------------------

                           *     *   * **   *  * *  *       *      * **
               222222222222222222222222222222222233333333333333333333333
               111111223344444455556666677777889900001112222222233333344
               004469283601113901141688904888588885689179124678923334594 7
               252696994702483681219145675145701680417783804101406921181
Consensus      TTCGGGGGCGCGTTATACTGGCACTACGCCGTCCCTCGCCGTCCCCATCCACCAAGC
JR436          ..TA..A.T......CCTG......G..........A.A...T...C..........
GB050          .....A.........CCTG.A....G......TT..G.A....AT..C.TGTA....
JR048          .C.......A.C.CT.....A....G.A..A...T.GT.......A.C.......A.
GB336          G..........C.C.....A.AGAACT....C...GG......A...C.........
GB361          ....A..CTA.CC...............AG.........AA.C.A..........G...
GB087          .C.......T........A.A....................A.......T.GT..G.A
GB369          .C.......T........A.A....................A.......T.GT..G.A
               ---][--------------------3-----------------------------------

                *        **           * ** *    *              *    *
               333333333333333333333444444444444444444445555555555555555
               445556666777778888999011233445556777788990002333345566668
               795570148344790169077346344333676028927010183011862345891
               412393224503751041389895089129085453757109557028852726160
Consensus      GAACTGCGCCCCTCGAGACGCGCAGAGGCGGGCCTGGGCGCCCCCGACCCTGGGGGC
JR436          A.G..A......C....G......A.A..A...........AA....T.........T
GB050          .G..C.T.T....T...G....G........TA.C.T..T.AGT..TGA...A....
JR048          .......T.T.....C..G....C...............T..AG..............
GB336          ...T.....T.G......GAA....G.......AC.......................T
GB361          ......T..AG....C............CG......C............T.T.A.A...
GB087          ...T...........T.A..AAT.........C....C.T.T....TA.G......AT.
GB369          ......T.......T.A..AAT..A.....C....C.T.T..G.TA.G..C...AT.
               ---------------------------------][4][---------5---------

                *     *             *
               5556666666666666 6666
               8890001122222222 2223
               2491225912335566 7780
               2880352628672409 5880
Consensus      TTCTGACGTTGGTCC  TCCG
JR436          .....G.T.............
GB050          .....T.........i1AA..
JR048          C.GA.G......A......G.
GB336          .C......CCTAAT...A...
GB361          .............A.T....G.
GB087          ....A.T.............C
GB369          ....A.T.............C
               -][-6][7][---8----][9]
```

FIG. 2.—Nucleotide polymorphism at *Xdh* locus in *Drosophila pseudoobscura*. The base-pair numbering of each polymorphic site is given, according to the numbering system of Riley (1989). A consensus sequence for the polymorphic sites is shown in boldface in the fifth row. Insertions and deletions are noted by i and d, respectively, and include the length of the insertion/deletion event, relative to the consensus sequence. The bottom row indicates the functional region for each polymorphic site. These regions are numbered as follows: 1 = 5′ flanking sequence; 2 = *Xdh* intron 1; 3 = exon 2; 4 = intron 2; 5 = exon 3; 6 = intron 4; 7 = exon 4; 8 = intron 5; and 9 = exon 5. The nonsynonymous polymorphisms are shown in boldface and are indicated by a star in the first row.

Synonymous polymorphisms are also not significantly clustered in their distribution ($G = 6.6$, df = 4, $P > 0.1$, by exons; $G = 10.68$, df = 5, $P > 0.05$, by blocks).

As expected, functionally distinct nucleotide sites at *Xdh* segregate very different

Table 2
Nucleotide Polymorphism at *Xdh* Region, Based on DNA Sequence Determination from Seven Isolates of *Drosophila pseudoobscura*

Region	No. of Base Pairs	No. of Polymorphic Sites	% Polymorphism	Nucleotide Diversity
5′ Flanking	196	7	3.6	0.014
Coding:				
Exon 1:				
Synonymous[1]	11	0	0	0
Nonsynonymous[2]	43	0	0	0
Exon 2:				
Synonymous	632	66	10.0	0.041
Nonsynonymous	1,981	22	1.1	0.004
Exon 3:				
Synonymous	269	18	6.7	0.028
Nonsynonymous	874	3	0.3	0.001
Exon 4:				
Synonymous	42	2	4.8	0.008
Nonsynonymous	126	1	0.8	0.001
Exon 5:				
Synonymous	10	1	10.0	0.021
Nonsynonymous	38	1	2.6	0.006
Intron:				
1	1,024	43	4.2	0.015
2	62	3	4.8	0.014
3	69	5	7.7	0.025
4	67	8	11.0	0.041
3′ Flanking	12	0	0	0
Summary:				
Synonymous	964	87	9.0	0.036
Nonsynonymous	3,062	27	0.9	0.003
Introns	1,222	59	4.8	0.017
Flanking	208	7	3.4	0.014
Overall	5,456	180	3.3	0.012

levels of variation. Two genes chosen at random from this sample differ, on average, at 68 of the 5,464 sequenced base pairs. Synonymous positions are, by far, the most highly polymorphic, with a nucleotide diversity of 0.036 (table 2). A *G*-test comparing the level of polymorphism for flanking, synonymous, nonsynonymous, and intron sites reveals that, among the functional classes of sites, there is significant heterogeneity in polymorphism ($G = 154.58$, df = 3, $P \ll 0.001$). Three of the polymorphic synonymous sites exhibit two independent mutational events; that is, there are three segregating base pairs at each of these positions. Synonymous polymorphisms are heterogeneously distributed among synonymous codons, are analyzed by degeneracy class—i.e., the two-, three-, four-, and sixfold degeneracy classes of Li et al. (1985)—and are weighted by the frequency of each degeneracy class in the consensus sequence ($\chi^2 = 14.97$, df = 3, $P \ll 0.01$). Twofold-degenerate codons are less polymorphic than expected, while sixfold-degenerate codons are more polymorphic than expected. This test corrects for the number of alternative bases per degeneracy codon.

Allele configurations of synonymous, nonsynonymous, and intron polymorphisms are given in table 3. Forty-eight percent of synonymous sites have only one

Table 3
Allele Configurations at Polymorphic Nucleotide Sites of *Xdh*

Configuration Patterns[a]	No. of Replacement Polymorphisms	No. of Synonymous Polymorphisms	No. of Intron Polymorphisms
6,1	23	42	37
5,2	3	28	9
4,3	0	14	11
4,2,1	0	1	1
3,3,1	1	2	0

[a] The first number is the number of isolates possessing the consensus nucleotide; the following number(s) is the number of isolates possessing the same nonconsensus nucleotide.

sequence with a nonconsensus base pair. Sixty four percent of intron sites and 85% of replacement sites also have a single unique sequence.

Amino Acid Polymorphism

Twenty-six of the 1,342 amino acid sites are segregating in this sample of seven sequences. Two amino acids have been multiply replaced; that is, three amino acids are segregating. In one case, the same nucleotide site is segregating for three nucleotides that encode three different amino acids. In the second case, two nucleotide sites within the same codon are responsible for the segregating amino acids.

The variation observed at *Xdh* is not randomly distributed among amino acids. Nine (32%) of the 28 replacements are nonconservative (nc) in charge, on the basis of charge state at neutral pH. Relative to the consensus sequence, the nc changes include one neutral residue replaced by a negatively charged one, four neutral residues replaced by positively charged ones, two negatively charged residues replaced by neutral alternatives, and one negative charge replaced by a positive residue. If one assumes equiprobable single-step mutation frequencies, there are significantly (i.e., 2.3 times) more nc changes than expected ($\chi^2 = 7.29$, df = 1, $P < 0.01$).

The high level of protein polymorphism observed at *Xdh* is due primarily to the accumulation of unique mutations rather than to the shuffling of a smaller subset of amino acid polymorphisms (see tables 3 and 4). On average, 8.7 amino acids differ between each polypeptide sequence, the range being 0–13 (see tables 4 and 5). Each polypeptide is distinguished by an average of 3.3 unique amino acids, the range being 0–7. Twenty-four (86%) of the 28 segregating sites are unique in this sample. JR436 and GB50 share two amino acid polymorphisms, and they also share an additional seven synonymous nucleotide polymorphisms. Although five of the nine shared polymorphisms are found in the first block of coding sequence, this clustering is not significant ($\chi^2 = 10.48$, df = 5, $P > 0.05$) and is interspersed with unique synonymous polymorphism. There are 41 nucleotide polymorphisms that distinguish between these two sequences, distributed along their entire lengths. The other two shared amino acid polymorphisms are found in GB87 and GB369, the two representatives of the major allozyme class. This pair differs at a total of only nine nucleotide sites.

On the basis of the number of residue differences per sequence comparison (table 5), certain sequences share a more recent common ancestor than do others. The recombination level inferred on the basis of four-cutter data (Riley et al. 1989) argues that there is no one phylogeny for the *Xdh* region. High levels of recombination make phylogenetic analyses, e.g., distance- or parsimony-based algorithms, inappropriate. However, the patterns of shared polymorphisms argue that there are two pairs of

Table 4
Polymorphic Amino Acids at *Xdh*

	Polymorphic Amino Acids at Nucleotide Position Number																									
	2414	2512	2684	2696	2707	2785	2981	3058	3091	3271	3342	3391	3448	3491	3740	3743	4165	4348	4349	4432	4704	5388	5656	5848	6196	6300
Consensus sequence	L	S	D	V	T	P	E	P	L	A	T	H	I	A	E	S	T	N	S	L	S	L	R	M	C	E
JR48	.	.	.	.	A	.	K	.	F	S	K	.	.	T	.	.	.	H	.	.	.	.	.	.	.	.
GB336	.	.	G	E	P	.	.	.	.	.	.	.	.	.	.	.	R	.	G	.	R	.	.	T	.	.
GB50	.	A	.	.	A	.	.	L	.	.	.	Q	.	.	G	.	.	.	.	.	.	M	.	.	.	.
GB369	.	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	D
GB87	.	.	.	.	.	.	.	.	.	.	.	.	V	.	.	.	.	.	.	.	.	.	.	.	.	D
JR436	.	A	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	F	.
GB361	P	.	.	.	.	A	.	.	.	.	.	.	.	.	.	C	.	.	.	V	.	.	H	.	.	.
Drosophila melanogaster	L	A	—	E	K	D	E	K	L	S	L	H	I	A	E	S	S	N	T	L	K	L	R	M	C	P

NOTE.—Save for GB369 and GB87, which belong to the same class, each line is a separate electrophoretic mobility class.

Table 5
Polymorphism Matrix: Nonsynonymous and Synonymous Differences between Each Pair of *Xdh* Sequences

	JR048	GB336	GB050	GB087	GB369	JR436	GB361
JR048		23	37	44	40	34	41
GB336 ...	13		41	30	36	36	33
GB050 ...	11	12		44	43	35	41
GB087 ...	9	9	8		5	42	39
GB369 ...	9	9	8	0		43	40
JR436	9	10	6	6	6		34
GB361 ...	12	12	11	7	7	9	

NOTE.—Data above the diagonal are number of synonymous differences; data below the diagonal are number of nonsynonymous differences.

sequences—i.e., the major electromorph pair (GB369 and GB87) and a minor and rare electromorph pair (JR436 and GB50)—that are more similar than others.

Table 4 provides a comparison between the polymorphic residues in *D. pseudoobscura* and the homologous sites from a representative *D. melanogaster* sequence. Nine of the polymorphic amino acids in *D. pseudoobscura,* including both amino acid sites segregating three residues in *D. pseudoobscura,* are either substituted or deleted in *D. melanogaster.* Three of the 28 sites compared differ from those in the *D. pseudoobscura* consensus—but are the same amino acid as the minor residue in *D. pseudoobscura.*

Discussion

High Levels of Amino Acid Variation at *Xdh*

Electrophoretic surveys have previously suggested high levels of variation for XDH (Singh et al. 1976; Keith et al. 1985). However, the sequence data reveal not only that there are many alleles of XDH but that variants can differ by as many as 13 amino acids. The allozymes chosen for the present study represent the range of electrophoretic mobility classes described by Keith et al. (1985). Thus it is likely that additional sequences of the remaining classes will not increase substantially the average of nine amino acid differences observed between allozymes.

It has been suggested that larger enzymes may tolerate more amino acid replacements that are effectively neutral in their physiological effect than can smaller proteins. XDH, with a subunit molecular weight of 146,898, is 2.5 times the size of esterase 6 (EST6) (Keith et al. 1985; Brady et al. 1990) and segregates an average of 2.5 times the number of amino acid differences between allozyme classes (Cooke and Oakeshott 1989). Of course, the tolerance of neutral substitutions will depend not only on size but on enzymatic function and evolutionary history of the enzymes examined. The enzymes ADH and EST6 are quite similar in size, yet ADH is often found to segregate very few allozymes while EST6 is highly polymorphic in the same species (Kreitman 1983; Cooke and Oakeshott 1989).

Constraints of Purifying Selection

It is possible to estimate the degree of purifying selection—or, rather, the level of functional constraint—experienced by a protein sequence. This requires an assumption that, at the DNA level, synonymous positions, experiencing little selective constraint, accurately reflect the neutral level of polymorphism for a given coding region. At *Xdh*, 9% of synonymous sites are polymorphic. Thus, roughly 276 nonsynonymous sites (0.09 × 3,062 bp) would be expected to be segregating if the protein were under no functional constraint. Only 27 of the predicted 276 nonsynonymous mutations for this sample are observed. Therefore, although electrophoretic surveys of XDH suggest it to be one of the most highly polymorphic enzymes in *Drosophila* (Singh et al. 1976), purifying selection is clearly distinguishing the vast majority of the mutations occurring at nonsynonymous sites. Is the remaining high level of observed amino acid variation simply that which escapes detection by purifying selection, or are these various forms of XDH being actively maintained by selection?

The patterns of polymorphism and divergence at *Xdh* argue that there may be regions of the protein that are free to evolve at an accelerated rate relative to other, more constrained regions. Both polymorphic and divergent replacement sites are clustered in the 5′ end of coding exon 2, and there is a significant reduction of both polymorphism and divergence in the 3′ end of the coding region. A *G*-test of independence reveals a significant coincidence between those codons that are substituted between *D. melanogaster* and *D. pseudoobscura* and those that are polymorphic within *D. pseudoobscura* ($G = 5.6$, df = 1, $P < 0.02$); that is, not only are certain regions of XDH under different levels of selective constraint, but it appears that certain codons are accumulating polymorphism and diverging at accelerated rates.

Tests of Neutrality

Neutral theory predicts that the level of polymorphism for a coding region should be positively correlated with the level of divergence, if the locus is evolving in a neutral fashion (Kimura 1983, pp. 25–33). A statistical test of that prediction, the HKA test (Hudson et al. 1987), has been applied to these data. The test compares the level of polymorphism segregating at two loci within a species to the level of divergence exhibited by those same loci examined between species. The data for this test are summarized in table 6. The coding region of *Xdh* was divided into two equal lengths, and these were treated as separate loci. The rationale for this grouping is based on a four-cutter survey of nucleotide polymorphism for this region, a survey that suggests that the sites are in linkage equilibrium and can thus be treated as if they are evolving independently (Riley et al. 1989). In addition, the 5′ end of the gene has been shown to exhibit both increased levels of polymorphism and divergence in reference to the 3′ end (Riley 1989). The HKA test was employed for both synonymous and nonsynonymous polymorphism separately. No significant deviations from the null hypothesis of neutrality were observed (synonymous—$\chi^2 = 0.95$, df = 1, $P > 0.3$; nonsynonymous—$\chi^2 = 0.004$, df = 1, $P > 0.95$).

Table 6
Number of Polymorphic and Divergent Positions at *Xdh*

	Xdh 5′ End		*Xdh* 3′ End	
	No. of Sites	No. of Differences	No. of Sites	No. of Differences
Polymorphism:				
Nonsynonymous . . .	1,510	17	1,510	10
Synonymous	503	58	503	29
Divergence:				
Nonsynonymous . . .	1,501	91	1,501	52
Synonymous	500	331	500	312

A quite different test of neutrality was developed by Tajima (1989), who examined the relationship between polymorphism and heterozygosity when different sized samples of restriction-site or sequence data are employed. Tajima produced a statistic, D, that compares an observed relationship between polymorphism and heterozygosity to predictions from neutral theory. We have applied Tajima's statistic to our total nucleotide polymorphism data for the *Xdh* region. The value of D for *Xdh* is 0.394. This value is not significantly different from 0, thus we cannot reject the null hypothesis of neutrality for the total nucleotide polymorphism at *Xdh.*

Both of these tests assume that a random sample is employed and that the populations are at stochastic equilibrium. The four-cutter data for *Xdh,* with their homogeneous distribution of restriction-site polymorphism both within and between allozyme classes, argue that even the stratified samples employed in our studies represent essentially random population samples (Riley et al. 1989). Several additional features of the four-cutter data, such as high haplotype diversity, linkage equilibrium, and large estimated population size, argue that the populations sampled are at stochastic equilibrium with respect to mutation and drift.

High Nucleotide Diversity Suggests Large, Stable Population Size for *D. pseudoobscura*

The allozyme classes included in the present study differ by 6–13 amino acids, with only two of the 28 polymorphic amino acids shared between the classes. This is a quite different distribution of amino acid polymorphism than was observed for ADH and EST6 (Kreitman 1983; Cooke and Oakeshott 1989). Allozymes of these latter enzymes can be linked via one or two amino acid differences. Allozymes of XDH are separated from each other by the accumulation of many mutational events. In addition, allele configurations for synonymous, nonsynonymous, and intron sites (table 3) reveal similar patterns in the accumulation of variation at these functionally distinct nucleotide positions. In other words, one does not need to invoke additional selective forces to explain either the level or the pattern of protein variation at *Xdh.*

The high level of nucleotide diversity exhibited by the sequences examined here argues both that population sizes of *D. pseudoobscura* are large and that they have remained large for a very long period of time. This interpretation is in agreement with results from the four-cutter restriction-map survey of the *Xdh* region, a survey which suggested that the effective population size for *D. pseudoobscura* is 2.4×10^6 (Riley et al. 1989).

Discriminatory Power of Sequential Protein Electrophoresis

Until quite recently, sequential polyacrylamide electrophoresis of proteins was the method of choice for characterizing the levels and patterns of genetic variation segregating in natural populations. Although it is clear that protein electrophoresis provides some measure of amino acid polymorphism, how accurate a measure these data provide remains unclear (Ramshaw et al. 1979; Coyne 1982; McClellan 1984). Our ignorance regarding the sensitivity of protein electrophoresis has led, in large part, to the failure of electrophoretic data to distinguish the importance of different evolutionary forces in maintaining protein polymorphism in nature (Lewontin 1985).

DNA sequence determination provides a complete description of amino acid variation, and such data can now be employed to reveal the resolving power of protein electrophoresis. DNA sequence determination of 11 *Adh* genes initially argued that protein electrophoresis was completely discriminating (Kreitman 1983). The two major

ADH allozymes revealed by protein electrophoresis were shown to differ by a single amino acid, and multiple representatives of the same allozyme class were shown to have identical amino acid sequences. A different result was obtained when 10 DNA sequences of the *Est-6* locus revealed one to seven residues differing between the encoded EST6 allozymes (Cooke and Oakeshott 1989). In addition, two representatives of each of two EST6 allozymes revealed three amino acids segregating within each class. However, this latter study employed high-resolution cellulose acetate electrophoresis, which may have different resolving powers than does sequential polyacrylamide electrophoresis. The results for XDH, which has 6–13 residues differing between allozyme classes, argues that there are no universal rules regarding the interpretation of electrophoretic data.

Attempts to use electrophoretic data to test evolutionary theory rely on electrophoretic mobility classes representing genotypes (Kimura and Crow 1964). In addition, most models—e.g., the charge state model introduced by Ohta and Kimura (1973)—assume that allozymes form ladders of single amino acid differences. The first assumption has been shown to be violated by EST6 in *D. melanogaster,* with three amino acid differences between representatives of the same allozymes being observed (Cooke and Oakeshott 1989). The second assumption is violated by both XDH and EST6, with high levels of variation—including multiple charge changes—being observed between allozyme classes. These violations lead to an underestimate of heterozygosity and, under certain models, would lead to a rejection of the null hypothesis of neutral evolution. For example, employing the Watterson test (Watterson 1974) which compares observed to predicted homozygosity under the infinite-alleles model, Keith et al. (1985) rejected the null hypothesis of neutral evolution. XDH deviated in the direction of too little diversity. In this case the rejection of neutrality may be due to the overestimate of genetic similarity inferred from electrophoretic data.

Conclusions

The data presented here reveal an unexpectedly high level of amino acid variation for XDH. Both the level and the pattern of this variation argue that population sizes for *Drosophila pseudoobscura* are quite large and that they have remained large for a very long period of time. The large population sizes ensure that mutations will occur frequently. The vast majority (>90%) of those mutations that occur at nonsynonymous sites can be distinguished by purifying selection and are rapidly eliminated from the population. Those amino acid polymorphisms remaining are neutral—or nearly neutral—in effect and segregate under the primary influence of random genetic drift. Allozyme classes that are not lost from the population by chance events slowly diverge from each other by the accumulation of additional neutral polymorphism. The longer the allozymes remain in the populations, the more distinct, in DNA sequence, they become.

The scenario just described does not explain the presence of the virtually identical sequences of the representatives of the major allozyme class that are included in the present study. If the allozyme classes of XDH are old, if nucleotide sites are in linkage equilibrium, and if population sizes are large, then the within- and between-allozyme class polymorphism are predicted to be fairly similar. Results from the four-cutter survey suggest that all three of these assumptions are valid and indicate similar within-allozyme-versus-between-allozyme class variation. However, with complete sequence information, we find virtual identity of DNA sequences for two representatives of the major allozyme class. How do we explain this discrepancy? Is the identity of the two sequences due to selection, or have we simply sampled two genes that have a recent common ancestry? The four-cutter data argue for the latter explanation; however, a definitive answer will require additional sequences of within-allozyme representatives. A small sample from each of two or three allozyme classes, particularly from the major class, should provide a resolution of this issue.

Acknowledgments

We thank Tim Keith for providing fly stocks and Lenard Diggins for maintaining them. We are grateful to Marty Kreitman, Paul Sharp, Charles Aquadro and, in particular, Richard Lewontin for their many helpful discussions and comments on the manuscript. We thank an anonymous reviewer for his or her many insightful comments on an earlier version of the manuscript. This work was supported by National Institutes of Health grant GM21179 to R. C. Lewontin and by a Sloan Fellowship to M.A.R.

LITERATURE CITED

BRADY, J. P., R. C. RICHMOND, and J. G. OAKESHOTT. 1990. Cloning of the esterase-5 locus from *Drosophila pseudoobscura* and comparison with its homologue in *D. melanogaster.* Mol. Biol. Evol. **7**:525–546.

BUCHANON, B. A., and D. L. E. JOHNSON. 1983. Hidden electrophoretic variation at the xanthine dehydrogenase locus in a natural population of *Drosophila melanogaster.* Genetics **104**:301–315.

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

COYNE, J. A. 1976. Lack of genetic similarity between two sibling species of *Drosophila* as revealed by varied techniques. Genetics **84**:593–607.

———. 1982. Gel electrophoresis and cryptic protein variation. Pp. 1–32 *in* M. C. RATTAZZI, J. G. SCANDALIOS, and G. S. WHITT, eds. Isozymes: current topics in biological and medical research. Vol. **6.** Alan R Liss, New York.

DENTE, L., G. CESARENI, and R. CORTESE. 1983. pEMBL: a new family of single stranded plasmids. Nucleic Acids Res. **11**:1645–1655.

HANAHAN, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166**:557–580.

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

KEITH, T. P. 1983. Frequency distribution of esterase-5 alleles in two populations of *Drosophila pseudoobscura.* Genetics **95**:467–475.

KEITH, T. P., L. D. BROOKS, R. C. LEWONTIN, J. C. MARTINEZ-CRUZADO, and D. L. RIGBY. 1985. Nearly identical allelic distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura.* Mol. Biol. Evol. **2**:206–216.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

KIMURA, M., and J. F. CROW. 1964. The number of alleles that can be maintained in a finite population. Genetics **49**:725–738.

KREITMAN, M. 1980. Assessment of variability within electromorphs of alcohol dehydrogenase in *Drosophila melanogaster.* Genetics **95**:467–475.

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

———. 1988. Molecular population genetics. Pp. 38–60 *in* L. PARTRIDGE and P. HARVEY, eds. Oxford surveys in evolutionary biology. Vol. **4.** Oxford University Press, Oxford.

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

———. 1986*b*. 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.

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

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MCCLELLAN, T. 1984. Molecular charge and electrophoretic mobility in cetacean myoglobins of known sequence. Biochem. Genet. **22**:181–200.

MANIATIS, T., E. F. FRITSCH, and J. SAMBROOK. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

OHTA, T., and M. KIMURA. 1973. A model appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–210.

RAMSHAW, J. A., J. A. COYNE, and R. C. LEWONTIN. 1979. The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics **93**:1019–1037.

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.

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

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

SINGH, R. S., R. C. LEWONTIN, and A. A. FELTON. 1976. Genetic heterogeneity within electrophoretic alleles of xanthine dehydrogenase in *Drosophila pseudoobscura.* Genetics **84**: 609–629.

TABOR, S., and C. C. RICHARDSON. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci. USA **84**:4767–4771.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:597–601.

WATTERSON, G. A. 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Prob. **6**:388–463.

BARRY G. HALL, reviewing editor

Received January 15, 1991; revision returned May 9, 1991

Accepted May 9, 1991