Origin and Evolution of a New Gene Descended From *alcohol dehydrogenase* in Drosophila

David J. Begun

Section of Evolution and Ecology, and Center for Population Biology, University of California, Davis, California 95616

Manuscript received August 10, 1996
Accepted for publication November 7, 1996

ABSTRACT

Drosophila alcohol dehydrogenase (*Adh*) is highly conserved in size, organization, and amino acid sequence. *Adh-ψ* was hypothesized to be a pseudogene derived from an *Adh* duplication in the *repleta* group of Drosophila; however, several results from molecular analyses of this gene conflict with currently held notions of molecular evolution. Perhaps the most difficult observations to reconcile with the pseudogene hypothesis are that the hypothetical replacement sites of *Adh-ψ* evolve only slightly more quickly than replacement sites of closely related, functional *Adh* genes, and that the replacement sites of the pseudogenes evolve considerably more slowly than neighboring silent sites. The data have been presented as a paradox that challenges our understanding of the mechanisms underlying DNA sequence divergence. Here I show that *Adh-ψ* is actually a new, functional gene recently descended from an *Adh* duplication. This descendant recruited ~60 new N-terminal amino acids, is considerably more basic than ADH, and is evolving at a faster rate than *Adh*. Furthermore, though the descendant is clearly functional, as inferred from molecular evolution and population genetic data, it retains no obvious ADH activity. This probably reflects functional divergence from its *Adh* ancestor.

PSEUDOGENES are, by definition, free of the selective constraints associated with the production of functional proteins. Thus, molecular evolution data from such genes are looked to as our best source of information on the spectrum of strictly neutral mutations and on the rates and patterns of substitution of such mutations. The simple, qualitative predictions for pseudogene evolution are clear: such genes are expected to evolve more quickly than their functional homologues, are expected to tolerate frameshift mutations, and are expected to show roughly equal rates of silent substitution per site and replacement substitution per site.

A number of papers appearing over the last 10 years have reported that several species in the *repleta* group of Drosophila have three, tandemly arranged copies of *Adh*, two functional genes (*Adh-2* and *Adh-1*) and a pseudogene (*Adh-ψ*) (FISCHER and MANIATIS 1985; ATKINSON *et al.* 1988; MENOTTI-RAYMOND *et al.* 1991; SULLIVAN *et al.* 1994); *D. mettleri* is thought to have an *Adh* pseudogene and only one functional *Adh* gene (YUM *et al.* 1991). Several pieces of evidence have been used to support the idea that *Adh-ψ* is nonfunctional. First, the presence of large numbers of frameshift mutations showed that there is no long open reading frame (ORF) common to all species in the group (summarized in SULLIVAN *et al.* 1994). Second, for species clearly having three copies of *Adh* (SULLIVAN *et al.* 1994), only two *Adh* isozymes are present (BATTERHAM *et al.* 1984). Third, in the two species examined, *D. hydei* and *D. mulleri*, the putative *Adh* pseudogene is transcribed only in pupae and/or adults (FISCHER and MANIATIS 1985; SULLIVAN *et al.* 1994), an expression pattern different from that of other Drosophila *Adh* genes (SULLIVAN *et al.* 1990).

However, some aspects of the data are incompatible with a simple pseudogene hypothesis. Though putatively nonfunctional (at least in terms of protein-coding capacity), this gene has evolved only slightly more quickly than *Adh-2* and *Adh-1* (SULLIVAN *et al.* 1994). This, in and of itself, is not damning to the pseudogene hypothesis, as large stretches of noncoding DNA could have unknown function or unusually low mutation rates. More problematic, however, are two facts regarding rates of replacement and silent site substitution. First, hypothetical replacement sites of the pseudogenes evolve only slightly more quickly than replacement sites of the functional genes, *Adh-2* and *Adh-1*. Second, pseudogene silent sites evolve ~10-fold more quickly than pseudogene replacement sites (SULLIVAN *et al.* 1994). It is difficult, if not impossible, to imagine how a gene with no protein-coding capacity could evolve in this fashion. Perhaps equally problematic for the pseudogene hypothesis is the presence of a putative 254-amino acid long open reading frame in two of the seven *repleta* group species examined, *D. hydei* and *D. buzzatii* (SULLIVAN *et al.* 1994). The maintenance of such an ORF in a gene hypothesized to have been mutationally inactivated several million years ago is most unlikely. Other aspects of the data are also difficult to reconcile

Address for correspondence: David J. Begun, Department of Zoology, University of Texas, Austin, TX 78712.
E-mail: djbegun@mail.utexas.edu

with the pseudogene hypothesis (*e.g.*, intron-exon boundaries and splice junctions are conserved, and codon bias is maintained in most species). These data have been presented as a paradox that challenges our understanding of molecular evolution (SULLIVAN *et al.* 1994).

Here I present data and analyses showing that *Adh-ψ* is not a pseudogene, but rather is a gene of unknown function descended from an *Adh* duplication. This gene recruited a large number of new N-terminal amino acids and subsequently became much more basic than its *Adh* ancestor. The new amino acids have the properties of a mitochondrial targeting presequence and are correlated with rapid, adaptive evolution in the remainder of the gene.

MATERIALS AND METHODS

All flies except for *D. hydei* were obtained from the Drosophila Species Center at Bowling Green University. Inbred lines of *D. hydei* were derived from individual, inseminated females caught at the Wolfskill Orchard, Winters, CA in Fall 1994. PCR products were amplified from genomic DNA from single flies using published sequence data for primer design. These products were made single stranded (HIGUCHI and OCHMAN 1989) and manually sequenced using internal primers designed from published sequence. For *D. eohydei*, some internal sequencing primers were designed using data from a cloned PCR product that had been sequenced on an ABI 377 automated sequencer. 5′-RACE products were generated following instructions of manufacturer (Clonetech), cloned (TA vector, Invitrogen) and sequenced on an ABI 377 automated sequencer. At least four RACE clones were sequenced from each species. The sequence data from this paper can be found under the following GenBank accession numbers: *D. eohydei* (U76468), *D. hydei* lines 1–9 (U76469–U76477), *D. mulleri* lines −1371.0 to −1371.22 (U76478–U76483), 5′ end of *D. hydei*, *D. mulleri*, and *D. mojavensis* (U76484, U76485, and U76486, respectively), *D. buzzatii* lines −1291.1 to −1291.7 (U77607–U77610).

RESULTS AND DISCUSSION

The *D. hydei* ORF: An important piece of evidence used to support the notion that the gene upstream of *Adh-2* is nonfunctional in the *repleta* group is the presence of frameshift mutations in codon 2 in most of the species examined (SULLIVAN *et al.* 1994, p. 446). However, the authors mistakenly compared nonhomologous "codons" (Figure 1). An example from *hydei* will make the point. The first four codons of *Adh-2* in *D. hydei* are ATG GCA ATC GCT. The homologous bases of *hydei Adh-ψ* were thought to be ATG -CA ATC GCT (bases 1752–1762 of MENOTTI-RAYMOND *et al.* 1991); the data were originally interpreted as a one-base frameshift deletion of a G at the first base of codon 2. However, the ATG codon of *Adh-ψ* is not homologous to the ATG codon of *Adh-2*. The 12 bases of *Adh-ψ* that are homologous to the first four codons of *Adh-2* are GAT GCA ATC GCT (Figure 1). Rather than a frameshift in codon 2, there are (at least) three substitutions in codon 1 (ATG to GAT). A 254-amino acid long ORF in *hydei* (SULLIVAN *et al.* 1994) starts at base 1751 of MENOTTI-RAYMOND *et al.* (1991). The same incorrect alignment appears to have been used at codon 2 for all the *Adh-ψ* sequences (SULLIVAN *et al.* 1994 and Figure 1).

Recruitment of new amino acids in *D. hydei*: One might be suspicious of the above conclusion, insofar as implicit in it is the notion that GAT rather than the typical ATG is the first codon of *Adh-ψ*. Remarkably, however, *Adh-ψ* has recruited a large number of amino acids 5′ to the codon homologous to the initiation codon of other *Adh*. *Adh-ψ* was hypothesized to have a small exon 5′ of the exon homologous to the first protein-coding exon of other *Adh* genes (SULLIVAN *et al.* 1994). These authors claimed that there was no evidence for an ORF that included this exon, though the data were extremely limited. I used rapid amplification of cDNA ends (5′-RACE) to isolate the putative N-terminal exon (or exons) of *Adh-ψ* from *hydei* (Figure 2). Sequences from the RACE product bear no obvious similarity to genomic sequence from what was initially thought to be the 5′-flanking region of *Adh-ψ* (MENOTTI-RAYMOND *et al.* 1991), yet comparison of the RACE data to sequence data from genomic DNA shows that the new exon(s) is contiguous with base 1736 of MENOTTI-RAYMOND *et al.* (1991). This demonstrates that a putative consensus splice junction upstream of the *Adh-ψ* homologue of the *Adh* initiation codon (SULLIVAN *et al.* 1994 and Figure 1) is functional, and supports the conclusion from Southerns that the new exon(s) is separated from the remainder of the *Adh-ψ* gene by a large intron (SULLIVAN *et al.* 1994). A methionine codon located downstream of a consensus Drosophila translation start (CAVENER 1987) defines the beginning of a 61-amino acid long ORF that is in frame with the 259-amino acid long ORF starting at base 1736 of MENOTTI-RAYMOND *et al.* (1991). The length of the new exon(s) plus untranslated leader is 251 bases, consistent with a previously published estimate of 252 bases from primer extension experiments (MENOTTI-RAYMOND *et al.* 1991).

In summary, a 320-amino acid-long open reading frame spans *Adh-ψ* in *D. hydei*, whereas other Drosophila *Adh* proteins are either 256 (*melanogaster* subgroup) or 254 amino acids (other Drosophila). The additional amino acids come from two sources, the new exon(s) (Figure 2) and five residues upstream of the homologue of the initiation codon of Drosophila *Adh* (Figure 1).

ORFs and new amino acids in other *repleta* group species: The fact that *hydei Adh-ψ* encodes a long ORF does not speak to the interpretation of the *Adh-ψ* data from other members of the *repleta* group. For example, all potential reading frames in *mojavensis* have multiple premature termination codons (SULLIVAN *et al.* 1994). Alignment of the protein-coding regions from published sequences of *mojavensis, mulleri, peninsularis* and *mettleri* to the correct reading frame of *hydei Adh-ψ* revealed the presence of frameshift mutations in each species relative to *hydei*. The long ORF homologous to that of *hydei-ψ* was destroyed by a single nucleotide

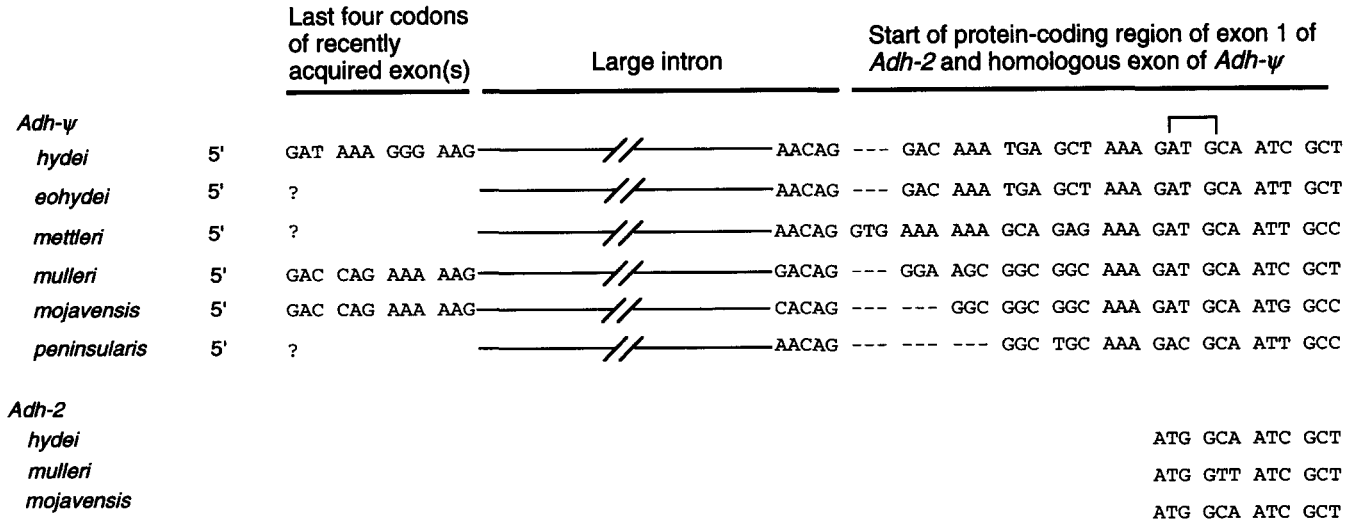

FIGURE 1.—Organization of the N-terminal portion of *Adh-ψ* in the *repleta* group and comparison to *Adh-2*. For *hydei, mulleri,* and *mojavensis* the last four codons of the new N-terminal exon(s) as inferred from comparison of RACE-derived and genomic sequence are shown. The large intron that follows is indicated by a horizontal line except for the last five bases. The first four codons of exon 1 from *Adh-2* from *hydei, mulleri* and *mojavensis* are aligned to the homologous codons of exon 2 from *Adh-ψ*. *Adh-ψ* data from species other than *eohydei* were previously published (SULLIVAN *et al.* 1994) or confirmed by sequence data collected for this study. The bracket shows the three bases that were previously thought to be the initiation codon of *Adh-ψ*. Alignment of the residues immediately upstream of the "GAT" *Adh-ψ* homologue of the *Adh* initiation codon are not meant to represent the only possible alignment. An insertion of a GCA codon between the eighth and ninth codons of the figure for *mettleri-ψ* is not shown. The 3′ intron splice junctions upstream of exon 2 of *Adh-ψ* in *hydei, mulleri,* and *mojavensis* were determined by comparison of RACE-derived and genomic sequence. The splice junctions for the remaining species were indirectly inferred based on homology and published sequences.

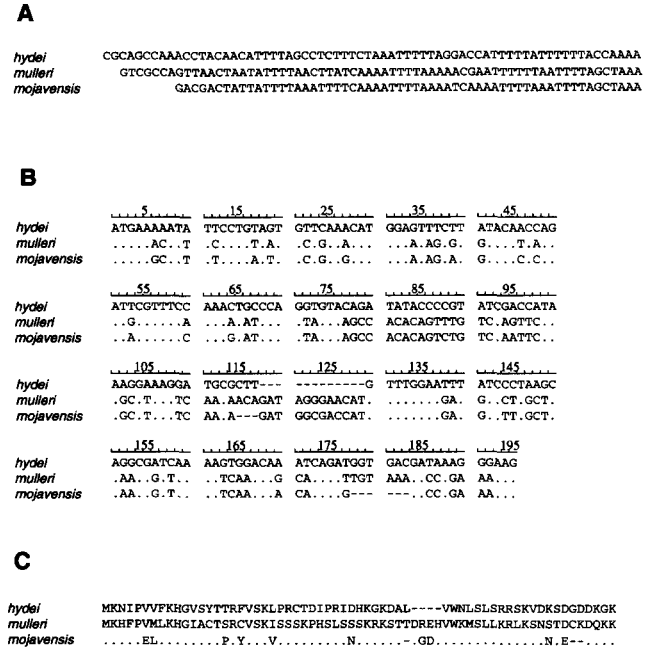

FIGURE 2.—New N-terminal exon(s) of *D. hydei, D. mulleri,* and *D. mojavensis*. (A) Untranslated leader. (B) Aligned protein-coding region. Conserved bases, ·; deletions, –. (C) Aligned predicted amino acid sequence. *D. mojavensis* residues identical to *D. mulleri* are indicated by ·. Deletions, –. For each species at least four independent RACE clones were sequenced. For the protein-coding region the consensus is shown; the longest untranslated leader is shown. The only size variation among clones within species was in the untranslated leader.

deletion in *mettleri* and a single nucleotide insertion in *peninsularis*. The *D. mulleri Adh-ψ* gene (FISCHER and MANIATIS 1985) had 14 single-bp insertions or deletions and a 27-bp deletion compared to *hydei*. *D. mojavensis* showed six single-bp insertions or deletions relative to *hydei*. My sequence data from PCR products amplified from genomic DNA in *mojavensis, mulleri, peninsularis* and *mettleri* showed that all putative single-base frameshift mutations in these species are attributable to previous laboratory error. The putative 27-bp deletion in *mulleri* resulted from incorrect assignment of an intron splice junction. These errors are described in Table 1. Thus, no special explanations are required to account for patterns of sequence conservation in the gene (SULLIVAN *et al.* 1994). I cannot speak to the issue of whether or not an intact copy of *Adh-ψ* is present in *D. mercatorum,* though it does seem unlikely that the large deletions reported to exist in the protein-coding regions amplified by SULLIVAN *et al.* (1994) could result from laboratory error.

The same procedures used to isolate the 5′ end of the *hydei Adh-ψ* gene were used to determine whether a homologous region was present in *mulleri* and *mojavensis,* two species from the *mulleri* subgroup of the *repleta* group (Figure 2). The *mulleri* and *mojavensis* exons are 65 and 62 amino acids long, respectively, and are identical at 52 of 62 alignable residues. Comparison of the sequences from the RACE reactions to those from genomic DNA shows that the *mulleri* exon(s) is contiguous with base 541 of FISCHER and MANIATIS (1985) (Figure 1), and that the *mojavensis* exon(s) is contiguous

TABLE 1

Changes that should be made to published *Adh-ψ* protein-coding regions

D. mojavensis (accession X12536)	*D. peninsularis* (accession L26039)
c insertion, 1267–1268	t deletion, 736
c insertion, 1366–1367	*D. mettleri* (accession M57300)
c insertion, 1378–1379	g insertion 1385–1386
a insertion, 1385–1386	
a insertion, 1388–1389	
g deletion, 1396	
D. mulleri (accession X03048)	
AG, 3′ splice junction is bases 702–703 (not 729–730)	
c deletion, 570	g insertion, 975–976
c insertion, 591–592	g insertion, 987–988
c insertion, 617–618	g insertion, 998–999
a insertion, 644–645	a insertion, 1009–1010
a deletion, 820	t insertion, 1011–1012
c insertion, 951–952	g insertion, 1068–1069
c insertion, 956–957	c insertion, 1322–1323
Bases 554–555 were CC in GenBank entry and were AA in all *mulleri* sequences determined for this paper.	

Numbering follows GenBank accessions. Intron positions of *mulleri Adh-ψ* were based on a comparison to functional *Adh* genes. By this method, the "first" intron of *Adh-ψ* was hypothesized to be 83 nucleotides long. This would be unusual in that this intron is considerably smaller in related species (55 bases in *mettleri, hydei* and *peninsularis;* 58 bases in *mojavensis*). In *mulleri* there is a potential 3′ intron splice junction (AG) at 702–703, consistent with an intron length of 57 bases. This is the correct 3′ splice junction and accounts for the putative 27-base deletion of the *Adh-ψ* protein in *mulleri. D. mettleri,* stock no. −1502.0; *D. peninsularis,* stock no. −1401.0; *D. mojavensis* and *D. mulleri* stocks given in Figure 3.

with base 1017 of ATKINSON *et al.* (1988) (Figure 1). The total length of the ORFs of the *Adh-ψ* proteins in *mulleri* and *mojavensis* are 324 and 320 amino acids, respectively. Given the phylogeny of the species discussed here (SULLIVAN *et al.* 1994; RUSSO *et al.* 1995) it is very likely that the organization of *Adh-ψ* is a shared, derived character of the *repleta* group.

Intraspecific variation: DNA sequence data from small population samples of *Adh-ψ* of *hydei, eohydei, mojavensis, mulleri,* and *buzzatii* (SCHAFER 1992) are shown in Figure 3 and Table 2. Each sequence was obtained from a single fly derived from an isofemale line. No frameshifts or premature termination codons are caused by any of the polymorphisms in any species. Silent site heterozygosity (WATTERSON 1975) is high in each species, at least an order of magnitude greater than replacement site heterozygosity, strongly supporting the conclusion that *Adh-ψ* codes for a functional, highly constrained protein. There is no convincing evidence of departures from the neutral model in MCDONALD and KREITMAN (1991) tests of the polymorphism and divergence data from *hydei* and *eohydei,* or *mulleri* and *mojavensis* (not shown). This could result from a lack of power (*e.g.,* only three alleles of *eohydei* were sequenced) or could reflect the possibility that much of the adaptive evolution in this gene occurred early in its history, before the split of the lineages leading to the species included in these analyses (see below).

Evolution of *Adh-ψ*: A database search revealed no similarity of the N-terminal exon to known proteins. This is not unexpected given the rapid rate of evolution of the new exon(s), even between the closely related species, *D. mulleri* and *D. mojavensis* (Figure 2). The amino acid composition of the *hydei* exon is highly biased toward positively charged (K or R) or hydroxylated (S, T, Y) residues, characteristics typically found in mitochondrial targeting presequences (HARTL *et al.* 1989). *D. mulleri* and *D. mojavensis* also have these features, despite the rapid rate of evolution of the primary sequence between these species and *D. hydei* (Figure 2). A previous analysis (JAUSSI 1995) showed that there is a strong trend for mitochondrial proteins to have a higher theoretical pI than their cytosolic homologues. Presumably this reflects adaptation for transport through the mitochondrial membrane and subsequent function in a higher pH environment (JAUSSI 1995). The mean theoretical pI of ADH-Ψ (9.22) is quite high compared to that of ADH (7.94); the distributions of pI for ADH-Ψ and ADH are almost nonoverlapping (Table 3). Thus, it appears that ADH-Ψ has become much more basic than its ADH ancestor, consistent with the hypothesis that the *Adh-ψ* protein is targeted to mitochondria. Future experiments should allow a rigorous test of this admitedly speculative idea.

In light of the radical change in the structure (Figure 2) and pI (Table 3) of the *Adh-ψ* protein compared to that of its highly conserved *Adh* ancestor, and given that the gene is clearly functional yet posseses no obvious ADH activity (see below), my working hypothesis is that positive selection has played an important role in the evolution of *Adh-ψ*. The *Adh-ψ* protein is evolving two- to fourfold more quickly than *Adh-2* in the *repleta* group

A) *D. hydei*

	1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
	8 0 0 0 1 2 2 2 3 3 3 3 3 3 3
	1 2 3 5 7 1 1 7 0 1 2 3 4 5
	6 4 9 1 7 3 9 3 0 6 4 1 8 5
1	T G C G A C A C G C T A G C
2	T G C G A C A C G C T A G C
3	C G C G A C G A A A T A G C
4	C G C G A C G A A A T A G C
5	C G C G A C G A A A T A G C
6	C G C A G A G C G C T A G T
7	C G T G A C A C G C T A G C
8	C A C G A C A A A A A T A C
9	C G C G A C G A A A T A G C

C) *D. mojavensis*

	1 1 1 1 1 1 1	1 1 1 1 1 1
	1 1 3 3 3 4 4	5 5 6 6 6 8
	5 6 2 8 8 4 8	4 9 4 4 7 5
	6 4 5 5 8 2 4	4 0 3 9 4 7
-1351.4	A ? C A G C G	T G T G T A
-1351.9	T T C A G C G	C/T G C/T G A/T G
-1351.12	A A T G G C G/T	C A T A T G
-1351.14	A A C A G T G/T	C G T G T G

B) *D. mulleri*

	0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
	6 6 6 7 0 0 0 0 0 1 1 1 1 1 1 1 2
	4 6 8 6 5 5 5 6 6 2 3 3 4 5 5 7 9 7
	0 1 2 9 4 7 9 0 3 6 0 3 3 4 5 3 3 7
-1371.0	G C A A T T G T C A A G A C A G T C
-1371.1	G C G A T T G T C T A A A C A A T T
-1371.16	G T A C C C C C C A T A G G T C A G C
-1371.18	G C A C C C C C C A T T G G T C A G C
-1371.19	A C A C C C C C C A T A G G T C A G C
-1371.22	G C A C C C C C C A T A G G T C A G C

D) *D. buzzatii*

	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	1 2 3 3 3 3 3 4 4 5 7 7 7 9 9
	9 0 1 2 3 4 4 1 5 9 5 8 8 3 9
	8 2 5 9 8 5 9 3 7 2 7 5 7 5 8
-1291.1	G C C C C G C G T G C T G T G
-1291.2	A A C C C G C A A G C G G T T
-1291.6	A C A C C C G A A G T T - T T
-1291.7	A C C G T G C G A A C T G C T

FIGURE 3.—Polymorphic sites in samples from natural populations of *repleta* group flies. Species Center line numbers for *mulleri, mojavensis,* and *buzzatii* are given to the left of nucleotide data. (A) Coordinates follow GenBank accession X58694. Silent variants are as follows: 1816, 2024, 2039, 2051, 2177, 2213, 2219, 2273, 2300. Remaining variants located in intron. (B) Coordinates follow uncorrected GenBank accession X03048. Sequences were from a single fly from each of six isofemale lines. Lines −1371.0, −1371.1, −1371.16, −1371.18, −1371.19, and −1371.22 were established from flies collected in Texas, Mexico, Cayman Islands, Conception Island, Haiti, and the Dominican Republic, respectively. Replacement variants are as follows: 640 (Gly/Ser), 1059 (Ser/Cys), 1173 (Ala/Thr). Silent variants are 769, 1054, 1057, 1060, 1063, 1193, 1277. Remaining variants are located in introns. (C) Coordinates follow the uncorrected GenBank accession X12536. Sequences were from a single fly from each of four isofemale lines originating from Baja, California. Three of the four flies were heterozygous at one or more bases. Nucleotide 1164 could not be scored in fly −1351.4 because of an indel; GenBank accession X12536 has bases 1162–1164 as "tta". The a is polymorphic in flies −1351.9, −1351.12, and −1351.14. In fly −1351.4 "tta" is replaced by "caaaaaaa." Polymorphism 1385 is really the base inserted between 1385 and 1386 of accession X12536. Replacement variant is 1649 (Lys/Arg). Silent variants are 1325, 1385, 1388, 1442, 1484, 1544, 1674, 1857. Remaining variants located in introns. (D) Coordinates follow GenBank accession no. U65746. Sequences were from a single fly from each of four isofemale lines. Lines −1291.1, −1291.2, −1291.6, and −1291.7 originated from Bolivia, Argentina, Argentina, and Australia, respectively. Replacement variants are 1198 (Asn/Ser) and 1413 (Gly/Ser). Silent variants are 1202, 1457 1592, 1757, 1935, 1998. Remaining variants are located in introns.

(SULLIVAN *et al.* 1994), a pattern consistent with the notion that the gene experienced functional divergence by means of accelerated amino acid substitution.

TABLE 2

Segregating sites and heterozygosity per nucleotide (WATTERSON 1975) in *Adh-ψ* genes

	Silent		Replacement	
	S	$\hat{\theta}$	S	$\hat{\theta}$
D. hydei (n = 9)	9	0.017	0	0.000
D. eohydei (n = 3)	2	0.008	0	0.000
D. mulleri (n = 6)	7	0.017	3	0.002
D. mojavensis (n = 7)	8	0.018	1	0.001
D. buzzatii (n = 4)	6	0.018	2	0.002

Each species was surveyed from the first base of the second exon (homologous to the first protein-coding exon of other *Adh*, see Figure 1) to the termination codon. There were two polymorphic sites in *eohydei* (2168, C/T; 2422, A/G). Both are silent with coordinates following those of *hydei*. The *eohydei* lines sequenced were −1631.0, −1631.1 and −1631.2 (Species Center, Bowling Green). S, segregating sites; $\hat{\theta}$, heterozygosity.

Further evidence for adaptation would come from the finding that the rate of evolution was more rapid during the early stages of the evolution of the new gene compared to the rate observed over more its more recent history. A nucleotide-based maximum likelihood approach (YANG 1995) was used to infer the DNA sequences at the internal nodes of the well-established tree in Figure 4 using only the 254 *Adh-ψ* codons homologous to *Adh* codons. Maximum likelihood approaches employing particular models of protein evolution (YANG *et al.* 1995) were deemed inappropriate for *Adh-ψ* because of assumptions or restrictions (*e.g.*, stationarity) that are incompatible with the data. The ratio of the inferred number of amino acid substitutions for branches AB *vs.* AC (22:2) is significantly different from the ratio of the number of changes below node B *vs.* below node C (56:27) (G-test; P = 0.02), supporting the idea that compared to ADH-2, the rate of evolution of the *Adh-ψ* protein was significantly greater early in its history compared to its more recent history. The fact that the theoretical pI changes from 8.43 to 8.84 along branch AB shows that evolution between reconstructed

TABLE 3

Theoretical pI of *Adh-ψ* and other *Adh* genes in Drosophila

	pI
hydei	
ψ exons 2–4	9.45
ψ exon 1	9.88
ψ exons 1–4	9.61
Adh-2	8.74
mojavensis	
ψ exons 2–4	9.01
ψ exon 1	10.15
ψ exons 1–4	9.50
Adh-2	8.47
mulleri	
ψ exons 2–4	8.73
ψ exon 1	10.31
ψ exons 1–4	9.49
Adh-2	7.73
buzzatii	
ψ exons 2–4	9.04
Adh-2	8.43
mettleri	
ψ exons 2–4	9.47
Adh-2	8.43
eohydei	
ψ exons 2–4	9.60
willistoni	8.77
immigrans	8.76
planitibia	8.44
pseudoobscura	8.43
subobscura	8.42
madeirensis	8.42
guanche	8.42
lebanonensis	7.80
nigra	7.75
wheeleri Adh-2	7.74
Z. tuberculatus	7.74
hawaiiensis	7.73
melanogaster	7.73
borealis	6.97
montana Adh-2	6.96
virilis Adh-1	6.52
ambigua	6.30

Calculated using the pI tool at the ExPASy WWW site (Geneva University Hospital and University of Geneva, Geneva, Switzerland). For *Adh-ψ* genes, exon 1 is the new N-terminal exon; exons 2–4 are homologous to protein-coding exons 1–3 of other Drosophila *Adh* genes.

hypothetical nucleotide ancestors has captured, at least qualitatively, some aspects of protein evolution known to have occurred along the lineage leading to the *Adh-ψ* protein (Table 3).

The observed sequences and reconstructed ancestral sequences were divided into two regions roughly thought to represent the cofactor binding domain (residues homologous to 1–140) and the substrate binding domain (residues 141–254) of *Adh* (BENYAJATI *et al.* 1981). For the first 140 residues the ratio of the number of amino acid sustitutions for branches AB *vs.* AC (12:2) is not significantly different from the ratio of the number of changes below node B *vs.* below node C (40:15) (G-test; $P = 0.33$). For the last 114 residues, AB *vs.* AC (10:0) is significantly different from below B *vs.* below C (16:12) (G-test, $P = 0.01$). Of these 10 replacement changes along branch AB, nine are charge-altering and eight are radical by the criteria of MIYATA *et al.* (1979). Only four of the 12 changes along AB for the first 140 residues were charge changes, and only four were radical by the MIYATA *et al.* (1979) criteria. The rate of replacement substitution per site for the final 114 residues in AB (0.042) is greater than the rate of silent substitution per site (0.030), a pattern often taken as *prima facie* evidence of strong positive selection for amino acid substitutions (*e.g.*, ENDO *et al.* 1996).

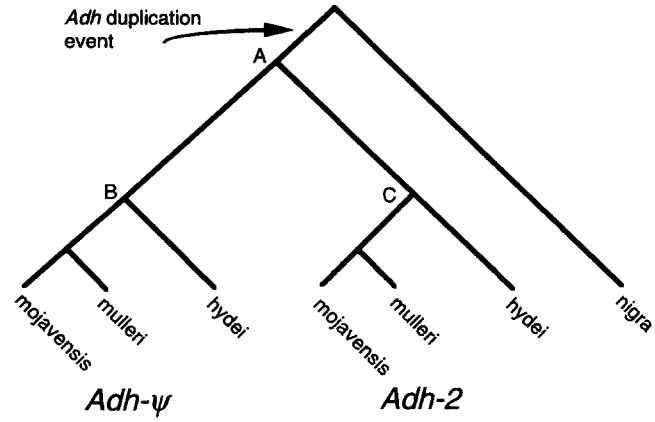

FIGURE 4.—Evolutionary relationship of *Adh-ψ* and *Adh-2* lineages. *D. nigra* is the outgroup. The time on the *Adh-ψ* side of the tree and the *Adh-2* side of the tree are the same. The model of nucleotide evolution in the likelihood analysis is from the DNAML progam of PHYLIP (FELSENSTEIN 1993), referred to as "F84" in YANG (1995). Comparison of branches AB and AC provides the best picture into the history of the genes before splitting of the *hydei* and *mulleri* subgroups.

There are several layers of uncertainty in these analyses, including error associated with reconstruction of ancestral sequences, doubt as to the locations of functional domains of ADH and as to whether such domains correspond to functional units in *Adh-ψ*, and uncertainty as to whether particular substitutions are radical with respect to function. Even so, the data taken as a whole support the idea that positive selection played a significant role in the early evolution of *Adh-ψ*, particularly in the carboxy-terminal third of the region homologous to *Adh*.

Though it was acknowledged that *D. hydei Adh-ψ* might conceivably have a long ORF homologous to that of *Adh*, SULLIVAN *et al.* (1994) argued that such a protein would be very similar to ADH (ADH-2 and ADH-Ψ of *D. hydei* are identical at 209 of 254 alignable residues) and so would almost surely have ADH activity; the absence of such activity was taken as supporting evidence for the pseudogene hypothesis (SULLIVAN *et al.* 1994). However, this conclusion may be unwarranted if functionally distinct proteins that are diverged at a small number of key residues still have high levels of overall

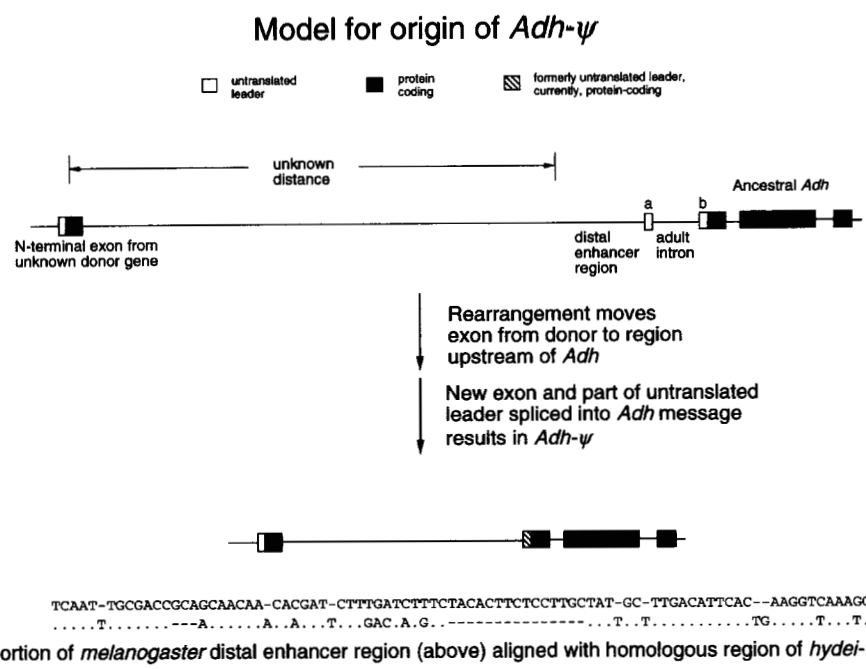

FIGURE 5.—The high sequence similarity of segments of the distal enhancer region of *melanogaster Adh* and the 3′ region of the large intron of *hydei Adh-ψ* supports the idea that the large *Adh-ψ* intron descended from the 5′-flanking region of an ancestral *Adh* with typical organization. The portion of the adult untranslated leader that was part of the first protein-coding exon of the ancestral *Adh* gene (b) became protein-coding sequence in *Adh-ψ*, while the remainder of the adult untranslated leader (a) was lost. The first base of the distal enhancer *melanogaster* sequence shown is −537 of Figure 2 of JEFFS *et al.* (1994). The first base of the *hydei* sequence is 1283 of MENOTTI-RAYMOND *et al.* (1991). –, deletions; ·, *hydei* bases identical to *melanogaster* bases.

similarity. Drosophila ADH from 26 species, including very distantly related taxa, were compared to amino acid sequences from *Adh-ψ* (*hydei-ψ, mulleri-ψ, mojavensis-ψ, buzzatii-ψ, peninsularis-ψ, eohydei-ψ, mettleri-ψ, buzzatii-2, hydei-2, mettleri, mojavensis-2, mulleri-2, wheeleri-2, immigrans, lebanonensis, pseudoobscura, willistoni, virilis-1, guanche, ambigua, borealis, flavomontana, grimshawi, lacicola, montana, Scaptomyza albovittata, planitibia, nigra, silvestris, heteroneura, yakuba, melanogaster, Zaprionus tuberculatus;* only the 254 residues alignable among all the sequences were used). Given the large amount of time in the genealogy of ADH for these species, we might infer that many, if not most, of the completely conserved residues cannot tolerate substitutions without the compromising of ADH activity. Therefore, fixed amino acid differerences between *Adh-ψ* and *Adh* are candidates for substitutions that modify the function of *Adh-ψ* so as to make its product unrecognizable by histochemical assays of ADH activity (BATTERHAM *et al.* 1984; SULLIVAN *et al.* 1994). There are three such residues, 43 (N in ADH to K in ADH-Ψ), 192 (H in ADH to N in ADH-Ψ), and 206 (E in ADH to Q in ADH-Ψ). The initiation codon also is a fixed difference (M in ADH to D in ADH-Ψ) (coordinates follow *D. melanogaster,* GenBank accession X60791). All four of these fixed amino acid differences were inferred to have occurred along branch AB in the aforementioned analysis of reconstructed ancestral nucleotide sequences. There are also a number of residues at which ADH and/or ADH-Ψ vary and at which the amino acids present do not overlap between genes. One might imagine that these are also possible candidates for substitutions associated with new function of *Adh-Ψ*. Finally, it is possible that lack of detectable ADH activity associated with the *Adh-ψ* protein results from interference of the recruited residues with domains required for such actvivity. Since we do not yet know the function of the *Adh-ψ* protein it is impossible to engage in informed speculation as to its ecological significance. I would only point out that all the flies discussed in this paper are capable of using rotting cactus cladodes as a substrate (summarized in BARKER and STARMER 1982; BARKER *et al.* 1990). One cannot help but wonder whether *Adh-ψ* is an adaptation for such a lifestyle.

Origin of *Adh-ψ*: At least two events, the recruitment of the new N-terminal exon(s) and the acquisition of new amino acid residues upstream of the ancestral ATG initiation codon of *Adh* (Figure 1) are required to explain the origin of *Adh-ψ*. There are currently no data that shed light on the source of the new N-terminal exon. For the moment we will posit that a genomic rearrangement (perhaps resulting from unequal crossing over) juxtaposed the first exon from an unknown donor gene to the 5′-flanking region of the ancestor of *Adh-ψ*. What subsequent events could account for proper splicing of this exon and for the new residues in exon 2 of *Adh-ψ*? The organization of *Adh* provides a plausible answer. *D. melanogaster Adh* has two transcripts, adult (initiated from a distal promoter) and larval (initiated from a proximal promoter). The adult transcript

includes an untranslated leader organized in two parts, separated by a 654-base intron (BENYAJATI *et al.* 1983 and Figure 5). This basic organization appears to be the ancestral condition in the genus (JEFFS *et al.* 1994; NURIMINSKY *et al.* 1996). Sequence similarity (85% over 65 alignable bases) between a segment of the large intron of *hydei Adh-ψ* and a segment of the 5′-flanking region of *Adh* in *melanogaster* (SULLIVAN *et al.* 1990) motivates a model for the origin of *Adh-ψ* (Figure 5). The new residues of exon 2 of *Adh-ψ* (Figure 1) are hypothesized to be remnants of the proximal portion of an ancestral, untranslated leader of *Adh*. The 3′ splice junction of the ancestral adult intron is hypothesized to have been co-opted as the 3′ splice junction of the new, large *Adh-ψ* intron. Thus, splicing of an untranslated exon in the ancestral *Adh* is hypothesized to have set the stage for both the recruitment of a new protein-coding exon and the conversion of an untranslated leader to protein-coding sequence in the descendant, *Adh-ψ*.

jingwei, a gene originally thought to be nonfunctional (JEFFS and ASHBURNER 1991), is also a recently evolved, chimeric Drosophila *Adh* (LONG and LANGLEY 1993). Though the molecular mechanisms creating *Adh-ψ* and *jingwei* were completely different (*jingwei* was created when a retrotransposed copy of *Adh* was inserted into the coding portion of another gene), both genes experienced rapid amino acid evolution and divergence in expression patterns (FISCHER and MANIATIS 1985; SULLIVAN *et al.* 1994) following recruitment of new N-terminal residues. This suggests two viewpoints (not necessarily mutually exclusive). First, recruitment of exons and functional divergence may be very common, even on relatively short evolutionary time scales. Second, selection on new function related to that of *Adh* may be very strong in Drosophila. Given that *Adh-ψ* is a functional gene, I propose that it be renamed *Adh-Finnegan* (Tim Finnegan, a character from an Irish folksong, was mistakenly declared dead and subsequently arose during his own wake).

I thank J. GILLESPIE for generous assistance in the maximum likelihood analysis, C. LANGLEY for typically insightful comments and discussion, and the Drosophila Species Center at Bowling Green University for stocks. Comments from A. G. CLARK and an anonymous reviewer are also appreciated. D. SCHAFER and J. S. F. BARKER kindly provided me with their *D. buzzatii* sequence. This work was supported by an Alfred P. Sloan Postdoctoral Fellowship in Molecular Evolution.

LITERATURE CITED

ATKINSON, P. W., L. E. MILLS, W. T. STARMER and D. T. SULLIVAN, 1988 Structure and evolution of the *Adh* genes of *Drosophila mojavensis*. Genetics **120:** 713–723.

BARKER, J. S. F., and W. T. STARMER, 1982 *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System*. Academic Press, New York.

BARKER, J. S. F., W. T. STARMER and R. J. MACINTYRE, 1990 *Ecological and Evolutionary Genetics of Drosophila*. Plenum Press, New York.

BATTERHAM, P., W. T. STARMER and D. T. SULLIVAN, 1984 Origin and expression of an *alcohol dehydrogenase* gene duplication in the genus Drosophila. Evolution **38:** 644–657.

BENYAJATI, C., A. R. PLACE, D. A. POWERS and W. SOFER, 1981 *Alcohol dehydrogenase* gene of *Drosophila melanogaster:* Relationship of intervening sequences to functional domains in the protein. Proc. Natl. Acad. Sci. USA **78:** 2717–2721.

BENYAJATI, C., N. SPOEREL, H. HAYMERLE and M. ASHBURNER, 1983 The messenger RNA for *alcohol dehydrogenase* in *Drosophila melanogasteer* differs in its 5′ end in different developmental stages. Cell **33:** 125–133.

CAVENER, D. R., 1987 Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Res. **15(4):** 1353–1361.

ENDO, T., K. IKEO and T. GOJOBORI, 1996 Large-scale searches for genes on which positive selection may operate. Mol. Biol. Evol **13:** 685–690.

FELSENSTEIN, J., 1993 Phylogenetic Inference Package (PHYLIP), University of Washington, Seattle.

FISCHER, J. A., and T. MANIATIS, 1985 Structure and transcription of the *Drosophila mulleri alcohol dehydrogenase* genes. Nucleic Acids Res. **13:** 6899–6917.

HARTL, F.-U., N. PFANNER, D. W. NICHOLSON and W. NEUPERT, 1989 Mitochondrial protein import. Biochim. Biophys. Acta **988:** 1–45.

HIGUCHI, R. G., and H. OCHMAN, 1989 Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17:** 5865.

JAUSSI, R., 1995 Homologous nuclear-encoded mitochondrial and cytosolic isoproteins. A review of structure, biosynthesis and genes. Eur. J. Biochem. **228:** 551–561.

JEFFS, P., and M. ASHBURNER, 1991 Processed pseudogenes in Drosophila. Proc. R. Soc. Lond. B **244:** 151–159.

JEFFS, P. S., E. C. HOLMES and M. ASHBURNER, 1994 The molecular evolution of the *alcohol dehydrogenase* and *alcohol dehydrogenase-related* genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **11:** 287–304.

LONG, M. Y., and C. H. LANGLEY, 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in Drosophila. Science **260:** 91–95.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptative protein evolution at the *Adh* locus of Drosophila. Nature **351:** 652–654.

MENOTTI-RAYMOND, M. A., W. T. STARMER and D. T. SULLIVAN, 1991 Characterization of the structure and evolution of the *Adh* region of *D. hydei*. Genetics **127:** 355–366.

MIYATA, T., S. MIYAZAWA and T. YASUNAGA, 1979 Two types of amino acid substitutions in protein evolution. J. Mol. Evol. **12:** 219–236.

NURIMINSKY, D. I., E. N. MORIYAMA, E. R. LOZOVSKAYA and D. L. HARTL, 1996 Molecular phylogeny and genome evolution in the *Drosophila virilis* species group: duplication of the *alcohol dehydrogenase* gene. Mol. Biol. Evol. **13:** 132–149.

SCHAFER, D., 1992 Developmental and molecular studies of *Adh* in *Drosophila buzzatii*. M. Sc. Thesis, University of New England, Armidale, New South Wales.

SULLIVAN, D. T., P. W. ATKINSON and W. T. STARMER, 1990 Molecular evolution of the *alcohol dehydrogenase* genes in the genus Drosophila. Evol. Biol. **24:** 107–147.

SULLIVAN, D. T., W. T. STARMER, S. W. CURTISS, M. MENOTTI-RAYMOND and J. YUM, 1994 Unusual molecular evolution of an *Adh* pseudogene in Drosophila. Mol. Biol. Evol. **11:** 443–458.

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

YANG, Z., 1995 Phylogenetic analysis by maximum likelihood (PAML), Version 1.1. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University.

YANG, Z., S. KUMAR and M. NEI, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641–1650.

YUM, J., W. T. STARMER and D. T. SULLIVAN, 1991 The structure of the *Adh* locus of *Drosophila mettleri:* an intermediate in the evolution of the *Adh* locus in the *repleta* group of Drosophila. Mol. Biol. Evol. **8:** 857–867.

Communicating editor: A. G. CLARK