Characterization of the Structure and Evolution of the Adh Region of Drosophila hydei

Marilyn Menotti-Raymond, W. T. Starmer and D. T. Sullivan

Department of Biology, Syracuse University, Syracuse, New York 13224

Manuscript received August 2, 1990
Accepted for publication October 5, 1990

ABSTRACT

Drosophila of the repleta group have a duplication of the gene which encodes alcohol dehydrogenase (ADH). We report the nucleotide sequence of an 8.4-kb region of genomic DNA of Drosophila hydei which includes the entire Adh region. Analysis of this sequence reveals similarity in organization to the Adh region of Drosophila mojavensis and Drosophila mulleri of the mulleri subgroup, with three genes ordered 5' to 3', Adh-ψ, Adh-2, Adh-1. Deletion of a nucleotide in the second codon of each pseudogene suggests that the first Adh duplication occurred before the divergence of the hydei and mulleri subgroups. However, Adh-1 and Adh-2 of D. hydei are significantly more alike than Adh-1 and Adh-2 of D. mojavensis. Models to account for the difference in similarity between the coding genes were tested by orthologous and paralogous comparisons of the extent of sequence divergence. A model which proposes that independent duplication events generated Adh-1 and Adh-2 in the two lineages is supported by these data. The D. hydei pseudogene is transcribed and the transcript is processed in a complex manner. An intron of greater than 6.2 kb exists between the first "coding" exon and an upstream exon which is approximately 250 nucleotides in length.

MEMBERS of species of the mulleri and hydei subgroups of the repleta group of Drosophila have been shown to contain a duplication of the gene which encodes alcohol dehydrogenase (ADH). Initially, the duplication was detected by the presence of isozymes that are differentially expressed during development (Oakeshott et al. 1982, Batterham, Starmer and Sullivan 1982). Subsequent molecular analysis of the Adh region of Drosophila mulleri (Fischer and Maniatis 1985) and of Drosophila mojavensis (Atkinson et al. 1988) have shown that these Adh loci contain a pseudogene, Adh-2 and Adh-1 arranged 5' to 3' over approximately 9 kb of DNA. Adh-1 and Adh-2 have different expression patterns during development. Adh-1 is expressed in early larval stages and Adh-2 is expressed in late third instar and adults. In addition, differences between species with respect to expression patterns have been observed (Batterham et al. 1984; Mills et al. 1986). Atkinson et al. (1988) have proposed a model for the evolution of the Adh region based on comparisons of nucleotide sequence divergence between the Adh genes of D. mojavensis and between the homologous genes of D. mojavensis and D. mulleri. They assumed that the ancestral Adh locus had a structure similar to the Adh locus of Drosophila melanogaster (Benyajati et al. 1983), i.e., a single gene with two promoters. The proximal promoter is used for larval expression and the other, for expression in late third instar larvae and in adults. The evolutionary model proposes that an initial gene duplication yielded two tandemly ar-

Marilyn Menotti-Raymond, W. T. Starmer and D. T. Sullivan

Department of Biology, Syracuse University, Syracuse, New York 13224

Manuscript received August 2, 1990
Accepted for publication October 5, 1990

ABSTRACT

Drosophila of the repleta group have a duplication of the gene which encodes alcohol dehydrogenase (ADH). We report the nucleotide sequence of an 8.4-kb region of genomic DNA of Drosophila hydei which includes the entire Adh region. Analysis of this sequence reveals similarity in organization to the Adh region of Drosophila mojavensis and Drosophila mulleri of the mulleri subgroup, with three genes ordered 5' to 3', Adh-ψ, Adh-2, Adh-1. Deletion of a nucleotide in the second codon of each pseudogene suggests that the first Adh duplication occurred before the divergence of the hydei and mulleri subgroups. However, Adh-1 and Adh-2 of D. hydei are significantly more alike than Adh-1 and Adh-2 of D. mojavensis. Models to account for the difference in similarity between the coding genes were tested by orthologous and paralogous comparisons of the extent of sequence divergence. A model which proposes that independent duplication events generated Adh-1 and Adh-2 in the two lineages is supported by these data. The D. hydei pseudogene is transcribed and the transcript is processed in a complex manner. An intron of greater than 6.2 kb exists between the first "coding" exon and an upstream exon which is approximately 250 nucleotides in length.

MEMBERS of species of the mulleri and hydei subgroups of the repleta group of Drosophila have been shown to contain a duplication of the gene which encodes alcohol dehydrogenase (ADH). Initially, the duplication was detected by the presence of isozymes that are differentially expressed during development (Oakeshott et al. 1982, Batterham, Starmer and Sullivan 1982). Subsequent molecular analysis of the Adh region of Drosophila mulleri (Fischer and Maniatis 1985) and of Drosophila mojavensis (Atkinson et al. 1988) have shown that these Adh loci contain a pseudogene, Adh-2 and Adh-1 arranged 5' to 3' over approximately 9 kb of DNA. Adh-1 and Adh-2 have different expression patterns during development. Adh-1 is expressed in early larval stages and Adh-2 is expressed in late third instar and adults. In addition, differences between species with respect to expression patterns have been observed (Batterham et al. 1984; Mills et al. 1986). Atkinson et al. (1988) have proposed a model for the evolution of the Adh region based on comparisons of nucleotide sequence divergence between the Adh genes of D. mojavensis and between the homologous genes of D. mojavensis and D. mulleri. They assumed that the ancestral Adh locus had a structure similar to the Adh locus of Drosophila melanogaster (Benyajati et al. 1983), i.e., a single gene with two promoters. The proximal promoter is used for larval expression and the other, for expression in late third instar larvae and in adults. The evolutionary model proposes that an initial gene duplication yielded two tandemly ar-
be expected of a species that had retained an ancestral Adh locus could represent an intermediate stage in the proposed model of ATKINSON et al. (1988). Conceivably D. hydei might have retained an Adh locus with one gene having two promoters and one gene having only a proximal promoter. The sequence of the Adh region of D. hydei reported here reveals this not to be the case. The locus is similar to the Adh regions of D. mulleri and D. mojavensis in many respects. Sequence comparisons of the Adh genes between and within these species suggest that the first Adh duplication event and the translational inactivation of the more 5' gene, generating a pseudogene, occurred prior to the divergence of the hydei and mulleri lineages. The second duplication events, however, probably occurred independently in the two lineages. We also report on several aspects of the structure and transcription of the Adh pseudogene which indicate an intriguing and not typical evolutionary history.

MATERIALS AND METHODS

Animals: The D. hydei strain used was collected by MARVIN WASSERMAN (Queens College) in Mexico and was polymorphic for electrophoretic alleles of Adh-1. Brother-sister pair matings were used to obtain a line homozygous for an electrophoretic allele of Adh-1. Morphology differences of the Adh-I allele that encoded an ADH-1 product with slower electrophoretic mobility than ADH-2.

Construction of D. hydei library: Genomic DNA, isolated according to the procedure of MILLS et al. (1986), was partially digested with Mbol and size-fractionated by centrifugation through a 5-29% NaCl gradient at 45K in a 50Ti rotor, 18° overnight. DNA fragments of 15-20 kb were isolated, treated with calf intestine alkaline phosphatase and ligated into BamHI digested phage arms of EMBL-4 (FRISCHAUF et al. 1983). The recombinant molecules were packaged using lambda packaging extract obtained from Amer sham. A library of 5 x 10^7 recombinant phage was amplified on Escherichia coli strain X559.

Isolation and characterization of Adh clones: A total of 120,000 recombinant phage were screened following the procedure of BENTON and DAVIS (1977) with nick-translated (MANIATIS, FRITSCH and SAMBROOK 1982) plasmid pLM19 (MILLS et al. 1986), which contains the coding region of the Adh gene of D. melanogaster. Probe was hybridized to phage DNA blotted onto nitrocellulose filters in 4 x SS C, 0.1% NaDodSO4, 0.1% sodium pyrophosphate (SPP), 10 x Denhardt's solution, (MANIATIS, FRITSCH and SAMBROOK 1982) and 50 ug/ml denatured salmon sperm at 68° overnight. Filters were washed in 4 x SS C, 0.1% NaDodSO4, 0.1% SPP at 68°, twice in 3 x SS C, 0.1% NaDodSO4, 0.1% SPP at 68° for 15 min, 2 x in 1 x SS C, 0.1% NaDodSO4, 0.1% SPP at 68° for 15 min and once in 0.5 x SS C at room temperature for 15 min. The DNA of the recombinant phage which hybridized to the Adh probe was mapped by restriction enzyme digestion patterns. Regions of the map with Adh similarity were determined by transfer of restriction fragments to nitrocellulose filters (SOUTHERN 1975) and hybridization to nick-translated pLM19 in 4 x SS C (MANIATIS, FRITSCH and SAMBROOK 1982), 1% NaDodSO4, 1 x Denhardt's, 100 ug/ml denatured salmon sperm at 68° overnight. Filters were washed as in library screening.

DNA sequencing: Several restriction fragments which spanned a 9.2-kb BglII-SalI region were subcloned into the m13 sequencing vectors mp18 and mp19 (NORRANGER, KEMPE and MESSING 1983). Nested sets of deletions were generated using the procedure of HENIKOFF (1984). Deletion clones were sequenced following procedures of SANGER, NICKLEN and COULSON (1977) using 25 SdATP (New England Nuclear) and separated on the Tris-HCl-boric acid-EDTA buffer gradient gels of BIGGIN, GIBSON and HONG (1983). Both strands of the 8.4-kb region were sequenced with the exception of 1.6 kb at the extreme 5' end of the sequence, a 1.1-kb region (positions 3698-4800, Figure 2), and 100 nucleotides at the extreme 3' end of the sequence. In these regions nucleotide sequence was determined using both the large fragment of DNA Polymerase 1 (Klenow fragment) (GIBCO BRL) and Sequenase (U.S. Biochemical). The 8.4-kb sequence was assembled using the programs of DNASTAR, Madison, Wisconsin.

Data analysis alignment: Comparisons of Adh genes both between and within species were by using the algorithm devised by WIBUR and LIPMAN (1983). Comparisons of introns and exons were as described by ATKINSON et al. (1988). Percent similarity between sequences was calculated as 100 X (number of nucleotides in common)/(total number of nucleotides compared). Alignments involving the pseudogene were accomplished by first adding a nucleotide at position 4 of exon 1 thereby compensating for the single frame shift in the D. hydei pseudogene.

Primer Extension and RNase Protection Assays: A 5.8-kb XbaI fragment including Adh-2 and the pseudogene was subcloned into XbaI-digested pBluescript SK+ (Stratagene) to yield plasmid pJDR. The pseudogene probe, pDPR, was constructed by ligation of a 284-bp Hael III fragment (positions 1555-1836, Figure 2) isolated from a XbaI-HindIII fragment of pJDR into Smal-digested pBluescript SK+. The pseudogene probe includes 85 bases downstream of the former ATG translation initiation site and 199 bases upstream. Complementary RNA probe labelled with 32PICTP (New England Nuclear) was generated using a Ribobasica II kit (IBI) and hybridized to 40 µg of pupal RNA at 45° overnight. RNase protection assays followed the procedure of ZINN, DIAIO and MANIATIS (1983). For the primer extension reaction 5' hydroxyl end labeling of oligonucleotides followed WOODS (1984). A 20 nucleotide synthetic oligonucleotide complementary to a region in the second "coding" exon of the pseudogene (positions 1913-1932, Figure 2) was hybridized to 10 µg of pupal poly(A)+ RNA at 45° for 5 hr and extended with AMV reverse transcriptase (Life Sciences) following the procedure of CALZONE, BRITTON and DAVIDSON (1987).

RNA sequencing: The 20-bp oligonucleotide used in primer extension analysis was end-labeled as for primer extension and hybridized at 45° for 5 hr to 10 µg of poly(A)+ RNA isolated from adult flies. RNA sequencing followed the chain termination procedure of GELIERTER (1987). The cDNA fragments were labeled with 32PdATP (New England Nuclear) and separated on the TBE buffer gradient gels of BIGGIN, GIBSON and HONG (1983).

RESULTS

A genomic library was screened with an Adh probe and two λ clones, RS5 and RS7, were obtained. Examination of restriction enzyme digestion patterns of phage DNA indicated that the clones consisted of overlapping regions of genomic DNA. Southern blots showed that clone λ RS7 contained 1.9- and 8-kb
Adh genes of D. hydei

D. hydei Adh region

The positions of restriction sites were determined using λRS7 DNA. The dark line below the map indicates the extent of genomic DNA that was sequenced. Shaded boxes indicate areas which hybridize to Adh coding region. EcoRI sites indicated at the extreme ends of the map are located in the EMBL-4 cloning junction. Direction of transcription is indicated above the map. The location of restriction sites for BgII (B), EcoRI (E), HindIII (H), SalI (S), XbaI (X), and XhoI (Xh) are shown. Arrows indicate restriction fragments which were subcloned into M13 vectors and sequenced; subclone 1: 2.4-kb; subclone 2: 2.1-kb SalI; subclone 3: 2.1-kb SalI; subclone 4: 2.3-kb SalI-EcoRI; subclone 5: 1.4-kb SalI-BglII; subclone 6: 0.6-kb EcoRI; subclone 7: 0.6-kb EcoRI; subclone 8: 1.9-kb EcoRI; subclone 9: 3.3-kb SalI-BglII.

FIGURE 1.—Restriction map of the Adh region of D. hydei. The positions of restriction sites were determined using λRS7 DNA. The dark line below the map indicates the extent of genomic DNA that was sequenced. Shaded boxes indicate areas which hybridize to Adh coding region. EcoRI sites indicated at the extreme ends of the map are located in the EMBL-4 cloning junction. Direction of transcription is indicated above the map. The location of restriction sites for BgII (B), EcoRI (E), HindIII (H), SalI (S), XbaI (X), and XhoI (Xh) are shown. Arrows indicate restriction fragments which were subcloned into M13 vectors and sequenced; subclone 1: 2.4-kb; subclone 2: 2.1-kb SalI; subclone 3: 2.1-kb SalI; subclone 4: 2.3-kb SalI-EcoRI; subclone 5: 1.4-kb SalI-BglII; subclone 6: 0.6-kb EcoRI; subclone 7: 0.6-kb EcoRI; subclone 8: 1.9-kb EcoRI; subclone 9: 3.3-kb SalI-BglII.

EcoRI fragments which hybridized to the Adh probe. It was concluded that the entire Adh locus of D. hydei was contained in a region of genomic DNA cloned in λRS7 since blots of genomic DNA had previously shown that the entire Adh region was contained on two EcoRI fragments of 1.9 kb and 8 kb (MILLS et al. 1986). Recombinant λRS5 contained a partial Adh region. Figure 1 is a detailed restriction map of λRS7 DNA. A set of restriction fragments spanning a 9-kb Bgl11-SalI region identified in Figure 1 were subcloned into M13 sequencing vectors and the nucleotide sequence of 8.4 kb was determined. The nucleotide sequence of this 8.4-kb region of D. hydei DNA, including the entire Adh region with a conceptualized translation, is shown in Figure 2.

In general, the organization of the D. hydei Adh region is similar to the Adh region of D. mulleri (FISCHER and MANIATI 1985) and D. mojavensis (ATKINSON et al. 1988). There are three Adh genes oriented 5' to 3' in tandem. The first two genes are separated by about 1 kb of DNA; the second two genes are about 2 kb apart. Each gene has the typical organization of a Drosophila Adh gene in respect to exon size, intron positions and polyadenylation site positioning. As observed in D. mojavensis and D. mulleri, the 5' gene is a pseudogene. The ATG codon of the pseudogene, analogous to an Adh translation start codon, is at nucleotide position 1752 of Figure 2. This gene is identifiable as a pseudogene by virtue of a single base pair deletion of the first nucleotide of the second codon which results in a frame-shift and resultant stop codon in the second exon. The Adh pseudogenes of D. mojavensis and D. mulleri have the same nucleotide deleted. The pseudogene has a sequence element, TATTTAA, at position 1567 which is identical to the distal promoter of D. melanogaster (BENYAJATI et al. 1983). The two downstream genes each have a sequence element, TATAAATA, at positions 3892 and 6589 identical to the proximal promoter of D. melanogaster (BENYAJATI et al. 1983). A sequence from 50 to 71 upstream of the Adh-2 TATA box and a corresponding sequence 53 to 73 bp upstream of the Adh-1 TATA box are highly similar to a corresponding region 45 to 62 nucleotides upstream of the proximal promoter of D. melanogaster. This region is conserved in all other Adh genes having a proximal-like promoter. In addition, the 5' flanking regions of Adh-1 and Adh-2 are very similar: alignment of 450 nucleotides upstream of the TATA boxes of the two genes reveals Adh-1 and Adh-2 are 73.2% alike. A high degree of sequence similarity has also been observed in corresponding regions of Adh-1 and Adh-2 of D. mulleri (FISCHER and MANIATI 1985) and D. mojavensis (ATKINSON et al. 1988) and has been used to argue that both coding genes are evolutionary derived from a gene having a proximal-like promoter. As observed in D. mojavensis and D. mulleri, the Adh-1 and Adh-2 upstream regions show greater similarity to each other than they do to sequence elements upstream of the pseudogene. However, approximately 400 bases upstream of the pseudogene ATG start codon, are sequences also found upstream of D. mojavensis and D. mulleri pseudogenes that have substantial sequence conservation and are very similar to a region upstream of the distal promoter of D. melanogaster (BENYAJATI et al. 1983). This region of conserved sequence is located in an area believed to act in an enhancer-like manner to confer distal expression on Adh-2 of D. mojavensis (ATKINSON et al. 1988) and to enhance the activity of the distal promoter of D. melanogaster (AYER and BENYAJATI 1990).

Each of the two downstream genes has Adh coding potential. The more 5' of the two (positions 3892 to 4776) encodes the more basic protein and thereby is judged to be Adh-2 as BATTERHAM et al. (1983) have shown that ADH-2, the adult isozyme, is the more basic of the two polypeptides in D. hydei. Therefore the most 3' gene (positions 6676 to 7560) encodes ADH-1. Only four amino acid differences are observed between ADH-1 and ADH-2. The coding regions of the exons are 97.3% similar and the introns are 94.0% similar. There are 27 nucleotide differences between these genes in a comparison of the aligned regions from the ATG translation initiation site to the stop codon. Sixteen of these substitutions occur at synonymous, or silent sites, which do not lead to an amino acid change. Four substitutions occur at nonsynonymous sites which do result in an amino acid change and seven substitutions occur in the introns.
**Adh genes of D. hydei**

**Table 1**

| Synonymous substitutions are nonrandomly distributed over the three exons (X^2 = 7.5, d.f. = 2, P < 0.05) with no substitutions observed in the first exon; five changes in the second exons and the majority of changes (11) found in the third exons. Substitutions in the introns were also nonrandomly distributed (X^2 = 4.36, d.f. = 1, P < 0.05) with the majority of substitutions observed in the first intron. Substitutions at three of the noncoding sites (one in exon two, two in exon three) result in no charge difference between the two polypeptides. A fourth nonsynonymous substitution, resulting in a neutral to basic residue change (glutamine to lysine) in exon three of **Adh-2**, accounts for the single charge difference between the two polypeptides.

In Table 1 we present comparisons of sequence divergence among the **Adh** genes of *D. hydei*. The extent of nucleotide substitution at synonymous sites, Ks, observed between **Adh-1** and **Adh-2** (Ks = 0.063, se = 0.024) is lower than the Ks observed between both **Adh-1** and **Adh-2** and the pseudogene (Ks = 0.272, se = 0.056 and Ks = 0.328, se = 0.067, respectively). Though not statistically significant, there is also an increase in the extent of substitution at nonsynonymous sites, Kd, measured in comparing either coding gene to the pseudogene as opposed to one another.

The difference in the amount of substitution for synonymous vs. nonsynonymous sites between the D. mojavensis pseudogene and either **Adh-1** or **Adh-2** was used to evaluate models proposed for the likely evolutionary history of the three **Adh** genes. Conclusions derived from a similar analysis of the difference in substitution for synonymous vs. nonsynonymous sites within the *D. hydei* **Adh** genes (data not shown) in general support the evolutionary model proposed for the *D. mojavensis** **Adh** genes. Additional information about the relationship of the *D. hydei** **Adh** genes can be gained by comparing the extent of
sequence divergence between the Adh genes of *D. hydei* with the extent of sequence divergence between the *D. mojavensis* Adh genes. As reported by Atkinson et al. (1988), the coding exons of Adh-1 and Adh-2 in *D. mojavensis* are 93.3% similar and the introns are 63.4% alike. The exons and introns of Adh-1 and Adh-2 of *D. hydei* are 97.3% and 94.0% alike, respectively. Therefore, the *D. hydei* genes are significantly more alike than Adh-1 and Adh-2 of *D. mojavensis*. We have considered two mechanisms which could account for the difference in similarity between Adh-1 and Adh-2 in *D. hydei* and *D. mojavensis*; the *D. hydei* genes are the products of a recent duplication event, independent of an older event that produced Adh-1 and Adh-2 in *D. mojavensis*. Adh-1 and Adh-2 of the two species are the products of a duplication event which occurred prior to the divergence of the lineage of the two species and subsequently a gene conversion event occurred between Adh-1 and Adh-2 of *D. hydei*. There are two ways to examine and distinguish between these possibilities. One is to do a phylogenetic analysis and construct a gene tree depicting the likely evolutionary history of the genes. Sullivan, Atkinson and Starmer (1990) have done this and their analysis supports the independent duplication model. The second method is to explicitly test the alternatives by comparing the salient differences that differentiate the two mechanisms. This was done by constructing four models of possible gene history as depicted in Figure 3. All models assume that the first Adh duplication occurred in a common ancestor of the two species. Model 1 assumes that the second duplication was generated independently. Model 2 assumes the second duplication occurred in a common ancestor of the two species. Models 3 and 4 are the same as Model 2 but include a gene conversion of either Adh-1 or Adh-2 in *D. hydei*. Each model predicts different levels of divergence of Adh genes within and between species (Figure 4) and a comparison of the observed patterns with the predictions can be used to evaluate the likely history of events in the evolution of the Adh locus of the two species. We determined which of the models is supported by comparisons of sequence divergence between Adh-1 and Adh-2 of *D. mojavensis* and *D. hydei*. We used two measures of sequence divergence: (1) % dissimilarity between aligned nucleotides of two Adh genes from ATG translation start to the termination codon and (2) \( K_\alpha \) (Table 2). All possible pairwise comparisons were made for each measure of divergence. In evaluating each pairwise comparison, we used a two by two G test (for evaluations of percent dissimilarity) or a \( t \) test (for evaluations of \( K_\alpha \)) to determine if differences in percent dissimilarity or \( K_\alpha \) observed between the two pairs in question were statistically significant. The two sets of paired comparisons of sequence divergence are presented in Figure 4. These data sets were compared with the four sets of predictions. Both sets of comparisons of sequence divergence support the predictions of model 1, which proposes independent second duplication events in the *D. mojavensis* and *D. hydei* lineages. This analysis does not include the possibility of multiple gene conversions. However, a double gene conversion, one in each species lineage would result in the same predictions (Figure 4) as given for model 1 (Figure 3). The ramifications and other evidence against this possibility are considered in the discussion. We also examined the 5′ untranslated and 3′ untranslated and flanking regions for paralogous (comparisons of the duplicate genes) and orthologous (between species comparison) sequence similarity. A comparison of the 3′ untranslated regions of Adh-1 and

---

**TABLE 1**

Nucleotide substitution comparisons of *D. hydei* Adh genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Adh-1</th>
<th></th>
<th></th>
<th>Adh-2</th>
<th></th>
<th></th>
<th>Adh-4</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%S(_E)</td>
<td>( K_\alpha )</td>
<td>( K_\beta )</td>
<td>%S(_I)</td>
<td>( K_\alpha )</td>
<td>( K_\beta )</td>
<td>%S(_I)</td>
<td>( K_\alpha )</td>
<td>( K_\beta )</td>
</tr>
<tr>
<td><em>D. hydei</em> Adh-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. hydei Adh-2</td>
<td>97.3</td>
<td>0.10</td>
<td>0.009</td>
<td>94.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. hydei Adh-(\phi)</td>
<td>81.1</td>
<td>0.78</td>
<td>0.102</td>
<td>74.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.106)</td>
<td>(0.065)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. mojavensis</em> Adh-1</td>
<td>88.4</td>
<td>0.48</td>
<td>0.044</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.039)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. mojavensis</em> Adh-2</td>
<td>88.9</td>
<td>0.44</td>
<td>0.042</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.044)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. mojavensis</em> Adh-(\phi)</td>
<td>82.8</td>
<td>0.58</td>
<td>0.106</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.078)</td>
<td>(0.070)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percent \( S_\alpha \) = percent similarity of exons; \( K_\alpha \) = substitution per nucleotide in introns; \( K_\beta \) = substitution per nucleotide for nonsynonymous sites; \( K_\gamma \) = substitution per nucleotide for synonymous sites; \( S_\gamma \) = percent similarity of introns; ND = not determined, numbers in parentheses are standard error. Variances were estimated according to equations 20, 21, and 22 (Li, Wu and Qiu 1985); percent similarity of exons was calculated as 100 \( \times \) (number of nucleotides in common)/(total number of nucleotides compared).
Adh-2 of *D. hydei* revealed they are 67.9% similar. We attempted to compare the 3′ untranslated regions of the four coding genes of *D. mojavensis* and *D. hydei* to identify areas of sequence conservation but other than a putative polyadenylation signal (AATAAA) observed in all four genes, no area of sequence conservation was found among the 3′ untranslated regions of the coding genes of *D. mojavensis* or *D. hydei*. The 5′ untranslated regions of *Adh-1* and *Adh-2* of *D. hydei* were 70.2% similar. In orthologous comparisons of the 5′ untranslated regions, *Adh-1* of *D. mojavensis* and *D. hydei* are 71.1% alike and *Adh-2* of the two species are 84.5% alike.

**FISCHER and MANIATIS** (1985) have reported that the *D. mulleri* Adh pseudogene is transcribed and we have investigated whether the *D. hydei* pseudogene is also transcribed. Figure 5 shows the results of transcript mapping of the pseudogene. The results from RNase protection analysis are shown in panel A. The pseudogene probe used in the analysis (coordinates 1552 through 1836, Figure 2) included 85 bases downstream of the ATG “initiation” site and 199 bases upstream. A 101 bp. fragment (85 bases downstream of ATG and 16 bases upstream of the ATG) of probe was protected from RNase digestion by RNA isolated from pupae. As the 16-base region upstream of the ATG is unique to the pseudogene, protection of this region could only come from a transcript of the pseudogene. The 5′ end of the protected fragment, position 1736 in Figure 2, is 163 nucleotides downstream of the distal TATA sequence and 16 bases upstream of the ATG (position -16). As this is an unlikely position for transcript initiation, we investigated whether -16 delineated a splice boundary of the transcript. The five bases immediately upstream of -16 constitute a potential 3′ acceptor splice junction (MOUNT 1982). We analyzed the transcript by primer extension, (Figure 5, panel B) and a product of approximate 394 nucleotides was obtained. As represented in panel D, the cDNA obtained includes 34 bases of second “coding” exon, 108 bases of first exon to position -16 and approximately 250 bases upstream of -16. As primer extension analysis indicates that pseudogene transcript extends beyond -16, this position must be a splice junction of the transcript. RNA sequencing of the pseudogene transcript was conducted to confirm this observation. Analysis of the 160 bp of sequence obtained revealed that pseudogene transcript is colinear with the genomic sequence up to position -18, at which point similarity between RNA and genomic sequence is lost, Figure 6. A search of the 1.6 kb of DNA sequence upstream of -16 revealed no regions similar to the 70 bases of transcript beyond the splice junction.

RNase protection analysis was conducted to determine if the 5′ region of the pseudogene transcript originated from regions within XRS7. Two restriction fragments which included the entire 6.2-kb region of
Figure 4.—Extent of sequence divergence in paired comparisons of Adh-1 and Adh-2 genes of *D. hydei* and *D. mojavensis*. For each comparison, the extent of divergence of pair A (horizontal axis) should be compared to pair B (vertical axis). 1, *D. hydei* Adh-1:*D. hydei* Adh-2; 2, *D. hydei* Adh-1:*D. mojavensis* Adh-1; 3, *D. hydei* Adh-2:*D. mojavensis* Adh-2; 4, *D. mojavensis* Adh-1:*D. mojavensis* Adh-2; 5, *D. hydei* Adh-1:*D. mojavensis* Adh-2; 6, *D. hydei* Adh-2:*D. mojavensis* Adh-1. This comparison is dependent on the positions of independent duplications relative to each other; only one possibility is shown in the figure. All predictions, <, =, >, are possible. Observations 1: Comparisons of percent dissimilarity of two pairs of Adh genes. Per cent dissimilarity between sequences was calculated as 100 × (number of nucleotides different)/(total number of nucleotides compared). For each comparison, pair A (horizontal axis) is compared to pair B (vertical axis). Observations 2: Pairwise comparison of values.

Table 2

<table>
<thead>
<tr>
<th>Pair No.</th>
<th>Paired Adh genes</th>
<th>Percent dissimilarity</th>
<th>Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>D. hydei</em> Adh-1:<em>D. hydei</em> Adh-2</td>
<td>3</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td><em>D. hydei</em> Adh-1:<em>D. mojavensis</em> Adh-1</td>
<td>15</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td><em>D. hydei</em> Adh-2:<em>D. mojavensis</em> Adh-2</td>
<td>14</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td><em>D. mojavensis</em> Adh-1:<em>D. mojavensis</em> Adh-2</td>
<td>9</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td><em>D. hydei</em> Adh-1:<em>D. mojavensis</em> Adh-2</td>
<td>13</td>
<td>0.44</td>
</tr>
<tr>
<td>6</td>
<td><em>D. hydei</em> Adh-2:<em>D. mojavensis</em> Adh-1</td>
<td>15</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Percent dissimilarity between aligned Adh-1 and Adh-2 genes of *D. hydei* and *D. mojavensis*: percent dissimilarity was calculated as 100 × (number of nucleotides different)/(number of nucleotides aligned).

ARS7 upstream of -16 (a 2.7-kb SalI-EcoRI and a 2.5-kb EcoRI, Figure 2) were subcloned into transcription vectors. RNase protection analysis using these probes detected no transcript along the coding strand for a region 6.2 kb upstream of position -16. Transcript(s)

were detected which were complementary to the opposite strand in both restriction fragments in samples of RNA isolated from pupae and adults, (data not shown). Therefore there is transcription on the opposite strand in the large intron which separates the 250 nucleotide upstream section of the transcript from the first "coding" exon of the pseudogene.

**DISCUSSION**

The structure of the Adh region of *D. hydei* is similar to the Adh regions of the previously analyzed species of the repleta group, *D. mulleri* (Fischer and Maniatis 1985) and *D. mojavensis* (Atkinson et al. 1988). The pattern of nucleotide substitution observed in comparing Adh-1 and Adh-2 is similar to that observed in other coding genes; a greater extent of substitution is observed in silent positions (synonymous sites and intronic positions) than at sites at which substitutions would result in an amino acid change (nonsynonymous sites). The distribution of synonymous substitutions
Adh genes of D. hydei

Figure 3.—Mapping of the D. hydei pseudogene transcript. Panel A, RNase protection assay. Lanes 1 and 2 show protection of complementary RNA transcribed from a 275-bp HaeIII fragment including 85 bases downstream of the former ATG initiation site and 190 bases upstream. Lanes: 1, pupal RNA; Lane 2, minus RNA control. The size marker shown was determined using an M13 sequencing ladder (not shown). Panel B, Primer extension analysis. A 20-mer oligonucleotide complementary to the second “coding” exon 2 (positions 1913–1932, Figure 2) was used. Lane 1, extension of pupal poly(A)+ RNA; the sequence shown includes transcript downstream of −16 colinear with genomic DNA, and transcript upstream of −18 which has no similarity to genomic DNA. Panel C, design of probe and its relation to the protected transcript fragment. Panel D, cartoon depicting primer extended product of ψ-gene. = exon; thin lines = flanking regions and introns; ■ = oligomer hybridized to second “coding” exon; darker horizontal lines = area of genomic DNA transcribed.

Among the three exons is nonrandom with the greatest number of substitutions occurring in the third exon. A similar observation has been made by Schaeffer and Aquadro (1987), with respect to the Adh genes of the melanogaster and obscura subgroup species. However, there is no such excess of synonymous substitutions in any exon of D. mojavensis (Menotti-Raymond, unpublished results).

The Adh loci of the repleta group have been derived from an Adh locus that was in all likelihood similar to the Adh locus that is currently found in all species which have a single Adh gene. Our goal has been to understand the series of events which occurred during the evolutionary transition from an Adh locus with a single gene to the complex structure that has been found in species of the mulleri subgroup. D. hydei is a member of a different subgroup than the previously analyzed species of the mulleri subgroup. We chose it for analysis in the hope that the lineage leading to D. hydei branched from the lineage leading to the mulleri species complex prior to some of the evolutionary events that have occurred in the Adh region of the mulleri species complex. If so, then it seemed possible that intermediates in Adh locus evolution might have been retained by D. hydei. It is apparent that at least two events occurred before the lineages of the hydei and mulleri subgroups diverged. The first was a duplication generating a locus with two tandemly arranged Adh genes. The second was mutational inactivation of the more 5’ gene resulting in the pseudo-gene. In support of this sequence of events is the observation that the first nucleotide in codon two of the upstream gene is deleted in each of the three species. This suggests a common origin of the pseudogene before divergence of the species lineages. Following these events it appears highly likely that the lineages leading to the mulleri and hydei subgroups diverged and a second duplication of the 3’ gene occurred in each lineage resulting in the two coding genes, Adh-1 and Adh-2.

This argument depends on several lines of evidence. The most important one is the conclusion that the higher sequence similarity of the two coding genes in D. hydei relative to D. mojavensis is due to an independent, more recent second duplication and not due to gene conversion. The incidence of gene conversion

Figure 6.—Sequence of D. hydei ψ-gene transcript. A 20-mer oligonucleotide complementary to the second “coding” exon 2 (positions 1913–1932, Figure 2) was used. Lane 1, extension of pupal poly(A)+ RNA; the sequence shown includes transcript downstream of −16 colinear with genomic DNA, and transcript upstream of −18 which has no similarity to genomic DNA.
between tandemly repeated genes is well documented. It has been shown to be instrumental in maintaining sequence homogeneity in members of the γ-globin locus (SLIGHTOM, BLECHL and SMITHIES 1980; POWERS and SMITHIES 1986), the heat shock genes of Drosophila (BROWN and ISH-HOROWICZ 1981) and the silk moth chorion genes (EICKBUSCH and BURKE 1986). FISCHER and MANIATIS (1985) report the only known example of a gene conversion in an Adh locus of the repleta group of Drosophila, a 200-bp region of gene conversion between Adh-1 and Adh-2 of D. mulleri. We considered the possibility of gene conversion in the D. hydei genes. However our analysis of paired comparisons of the extent of sequence divergence between the coding genes of D. hydei and D. mojavensis provides no evidence that gene conversion is the reason for the high similarity of the D. hydei genes. Rather, our analysis fully supports the alternative explanation. Two measures of sequence divergence, percent dissimilarity and $K$, suggest that D. hydei Adh-1 and Adh-2 are the products of a recent duplication event, independent of an older event that produced Adh-1 and Adh-2 of D. mojavensis.

The models we present in Figure 4 distinguish between independent second duplication events and a single gene conversion event. However, those models do not rule out multiple gene conversions occurring at different times as the cause of different similarities of the two coding genes in each lineage. It is possible that the first and second Adh gene duplications occurred in a lineage common to D. hydei and D. mojavensis. Subsequent to divergence of the two species lineages, separate gene conversion events occurred between Adh-1 and Adh-2 of each lineage. If the conversion events were widely separated in time, the consequences of independent gene conversion events could be similar to the consequences of independent duplication events. An approach to eliminating this possibility in the present case is to ask whether there is evidence for gene conversion in the lineage of D. mojavensis since this would be required under the multiple gene conversion hypothesis. ATKINSON et al. (1988) previously considered whether gene conversion had occurred in D. mojavensis Adh-1 and Adh-2 by comparing the extent of sequence divergence in orthologous and paralogous comparisons with D. mulleri Adh-1 and Adh-2. The extent of sequence divergence observed between D. mojavensis Adh-1 and Adh-2 in orthologous comparisons is similar to the amount of divergence between homologous D. mojavensis genes and D. mulleri genes. This suggests that Adh-1 and Adh-2 of D. mojavensis have diverged without gene conversion for approximately as long as the two species have diverged. A further attempt to detect gene conversion in Adh-1 and Adh-2 of D. mojavensis was made using the analysis of POWERS and SMITHIES (1986) to identify small regions of gene conversion in the human fetal globin genes. Eight Adh coding genes of the mulleri and hydei subgroups were aligned from the ATG translation initiation codon to the stop codon. These included Adh-1 and Adh-2 of D. mojavensis (ATKINSON et al. 1988), D. mulleri (FISCHER and MANIATIS 1985) and D. hydei, Adh of D. mettleri (J. YUM, W. T. STARMER and D. T. SULLIVAN submitted) and Adh-1 of D. navoaja (WEAVER, ANDREWS and SULLIVAN 1989). A position by position analysis over the entire contiguous stretch of DNA was used to identify positions where intraspecific similarity (duplicate genes within a species) were greater than interspecific similarity (homologous genes between species). If two consecutive positions were found with greater intraspecific similarity, a region of gene conversion was suggested. No significant regions of gene conversion were observed in D. mojavensis Adh-1 and Adh-2 using this analysis (data not shown). Therefore from analyses of models of the proposed evolution of the Adh region in D. hydei, and from the lack of evidence to suggest that gene conversion has been operative in the coding genes of D. hydei or D. mojavensis, it is likely that independent duplication events generated Adh-1 and Adh-2 in the mulleri and hydei lineages. Though the possibility of gene conversion cannot be formally eliminated, all the evidence is contrary to this possibility.

The outcome of gene conversion and duplication can be similar i.e., two tandemly repeated genes of high sequence similarity. However, the state immediately before these events is very different. In the simplest cases, gene conversion is initiated from two tandem genes while a duplication initiates from a single gene. Therefore, another way to distinguish between these alternatives is to determine the structure of the locus before the event in question by identifying a species having retained the ancestral structure, i.e., a species positioned in the same lineage but which has diverged before the event in question. Recently the structure of the Adh locus of D. mettleri has been found to contain a pseudogene and a single coding gene (J. YUM, W. T. STARMER and D. T. SULLIVAN, unpublished data). The existence of this two gene locus strongly suggests a recent origin for the second duplication event. More importantly, a gene tree of Adh genes calculated using % similarity or $K$, places the D. mettleri coding gene on the lineage to D. hydei following the divergence of the lineage to D. mojavensis and D. mulleri. This means that there were species with a single coding Adh gene in the hydei lineage subsequent to the divergence of the mulleri lineage.

The issue of distinguishing between gene conversion and independent duplications of the Adh genes is important because we know gene conversion is usually
a common process in tandemly repeated genes. This leads to the question as to why evidence of its occurrence is not more commonly found in the Adh genes of the repleta group. There seem to be two possibilities. Either there are fundamental differences in the genetic properties of this locus in these species or gene conversion occurs at a usual frequency but for some reason the products do not persist in the populations we have sampled. There is at present insufficient data to distinguish between these alternatives.

Pseudogenes have usually been thought to be regions free of selection (Li, Gojobori and Nei 1981). However, Atkinson et al. (1988) have reported a curious conservation of the D. mojavensis Adh pseudogene. In a comparison of nucleotide substitutions between the D. mulleri and D. mojavensis pseudogenes, exons were found to be 91.7% similar, closely approximating the similarity of the orthologous coding genes of Adh-1 (94.5%) and Adh-2 (94.4%) of the two species. More unexpectedly, the pseudogene introns are more highly conserved (84.2%) than are the introns of the orthologous coding genes. (The Adh-1 introns are 67.8% alike and the Adh-2 introns are 71.2% alike). The D. hydei pseudogene also shows a high degree of conservation relative to an Adh coding gene. Whereas it was not possible to make orthologous comparisons of the D. mulleri and D. mojavensis Adh introns as the amount of sequence divergence precluded effective alignment, the exons of the pseudogenes of these species are 80.1% similar. This suggests constraint on the divergence of the D. hydei pseudogene since both pseudogenes and introns are expected to diverge rapidly. Additionally, many of the characteristics of a functional Adh gene are retained by the D. hydei pseudogene, including the size of coding exons, positions of introns and putative polyadenylation site and acceptor and intron splice junction sequences. However, as Starmer and Sullivan (1989) report, the codon bias characteristic of Adh genes is lost in the D. hydei pseudogene; this is consistent with its loss of Adh coding function. Furthermore there is no substantial open reading frame that is common to the pseudogenes of D. hydei, D. mulleri and D. mojavensis. Nonetheless the D. hydei pseudogene is transcribed and has a complicated splicing pattern leaving open the possibility of some other function unrelated to translation.

This research was supported by U.S. Public Health Service grant GM-31857 to D.T.S. We thank KATHY WOJAS for isolating recombinant λARS7. RHONDA LEE for sequencing the 5′ end of the pseudogene transcript and RICH SOBEL for computer expertise.

LITERATURE CITED


Communicating editor: J. R. Powell