Unusual Molecular Evolution of an Adh Pseudogene in Drosophila

David T. Sullivan, William T. Starmer, Sandra W. Curtiss, Marilyn Menotti-Raymond, and Jungsum Yum

Department of Biology, Syracuse University

The Adh locus in Drosophila species which are members of the repleta group contains products of one or two duplication events. In all species examined to date one of the Adh genes is now a pseudogene, since mutations have rendered these genes incapable of being translated into a functional alcohol dehydrogenase. These pseudogenes contain introns in the standard Adh gene position; hence, their origin is not by retrotransposition. Comparison of the sequences of the Adh-Y from representatives of each of the subgroups of the repleta group reveal that the Adh pseudogene is present in each subgroup and that mutations at codon 2 and a deletion in the region immediately 5' to Adh-Y are common to all species. Therefore, it is likely that the translational inactivation event that resulted in a pseudogene occurred before the divergence of the species that make up the repleta group. We have investigated the transcription of Adh-Y of D. hydei and have found that the transcription has a developmental profile dissimilar from any known Adh gene, does not utilize an Adh promoter, and is initiated at a point almost 12 kb upstream. Comparison of sequence divergence of Adh-Y within species of the repleta group reveals that rates of evolution of the exons of Adh-Y are substantially slower than intergenic regions and are only slightly faster than those of exons of functional Adh genes. Second, retention of codon bias is found in the Adh-Y of most species, and substitution at synonymous coding positions substantially exceeds substitution at nonsynonymous coding positions. Comparison of the evolution of other putative pseudogenes with repleta group Adh pseudogenes suggests that at least some pseudogene sequences in Drosophila may be evolving through mechanisms and/or under influences not presently understood.

Introduction

The study of gene duplication in the genus Drosophila provides a unique opportunity for the understanding of molecular genetic events that occur during evolution. Gene duplications allow a series of orthologous and paralogous comparisons that provide a basis for detailed molecular evolutionary analyses, to uncover the important genetic changes that have occurred in structural and regulatory sequences during evolution. Comparisons can be made with assurance that differences between paralogous genes occur after the time of duplication, and differences among orthologous genes occur after the divergence of species from which the genes were isolated. Analyses of genes in the genus Drosophila can be based on the substantial body of knowledge about species phylogeny and ecological genetics of numerous species groups in the genus and can utilize a broad molecular genetic background.

Using this conceptual framework, we have conducted a series of studies on the molecular evolution of Adh genes that have undergone duplication during the evolution of species of the repleta group of the genus Drosophila (Batterham et al. 1984; Mills et al. 1986; Atkinson et al. 1988; Menotti-Raymond et al. 1991; Yum et al. 1991). These results and those of others enabled us to describe the sequence of events that have occurred during evolution of the multiple Adh genes found in species of the hydei and mulleri subgroups (Yum et al. 1991). The repleta group of the subgenus Drosophila radiated in the New World ~30 Mya (Throckmorton 1982). For the most part, primitive members of the repleta group probably lived in forest habitats on fallen fruits. Subsequently, diversification of the group was coincident with the spread of arid regions ~20 Mya and resulted in five major subgroups—repleta, mercatorum, hydei, mulleri, and fasciola (not shown; see fig. 1)—that now have members that use rotting stems, trunks, cladodes, and fruits of columnar and prickly pear cactus. Host specificity seems to be the rule
for these species, so that divergence in the cacti provided a framework for speciation in the drosophilids.

Studies on the Adh genes found in repleta group species show that most have three tandemly arranged genes (Adh-Ψ, Adh-2, and Adh-1) (Fischer and Maniatis 1985; Atkinson et al. 1988; Menotti-Raymond et al. 1991). These span about 9 kb of DNA. In D. mettleri, which has only two Adh genes, Adh-Ψ and Adh are also in tandem (Yum et al. 1991). The Adh pseudogenes (Adh-Ψ) are homologous to the original Adh gene of the lineage. Since Adh-Ψs have introns in the same positions as functional Adh genes, they have not been generated through retrotransposition. Adh genes found in Drosophila species that are not in the repleta group, such as D. melanogaster (Benyajati et al. 1983), have one promoter (the distal promoter), which is separated from the coding region by a large intron in the 5′ untranslated region, and a second promoter, which is close to the coding region (the proximal promoter). However, in species of the repleta group, only the proximal promoter of the Adh genes has remained after the duplications, so that each functional gene has only a single, proximal-type promoter (Yum et al. 1991).

A fundamental assumption about the molecular evolution of pseudogenes is that their nucleotide sequences are under no functional constraints, because pseudogenes can make no protein product, have no other function, or otherwise do not contribute to the phenotype (Li and Gojobori 1983). This idea has been investigated with respect to the evolution of globin pseudogenes and the accumulation of mutations that have rendered them translationally inoperable (Shapiro and Moshirfar 1989). Quantitative analysis of the goat β-globin pseudogene evolution, which like Adh-Ψ is not a retropseudogene, showed that nucleotide substitutions occurred in a random fashion with no polarity, that there were no differences with respect to coding and noncoding regions, and that there were no differences between synonymous and nonsynonymous sites (Shapiro and Moshirfar 1989). This is exactly the evolutionary history that is expected for sequences that do not encode a functional product.

Initial molecular evolutionary comparisons of Adh-Ψ in the closely related species D. mulleri and D. mojavensis revealed relatively more sequence similarity than might have been expected for pseudogenes (Atkinson et al. 1988). We speculated that an ancestor of Adh-Ψ might have been functional for the period just preceding divergence of the mulleri species complex, of which D. mulleri and D. mojavensis are members. Subsequently, we sequenced Adh-Ψ in the more distantly related species, D. hydei (Menotti-Raymond et al. 1991) and D. mettleri (Yum et al. 1991), and again found a relatively high degree of sequence conservation among this set. Consequently, we set out to determine more precisely the time of the inactivating events that led to the origin of Adh-Ψ and to evaluate its molecular biology and its evolution.

**Material and Methods**

DNA for sequencing of the Adh pseudogenes of Drosophila peninsularis and D. mercatorum was obtained by polymerase chain reaction (PCR) amplification from genomic DNA by using a 5′ primer whose sequence is located in the conserved upstream adult enhancer region (Ayer and Benyajati 1990) and a 3′ primer whose sequence is located in a region homologous to exon one in the downstream Adh-2 gene. PCR amplification with this primer set generates DNA fragments that extend from the enhancer region through the pseudogene and intergenic region to the downstream Adh gene. Amplified fragments were cloned into the vector pBluescript SK+ (Strategene) and sequenced using single-stranded sequencing and Sequenase (U.S. Biochemicals). fly stocks of these two species were obtained from the Drosophila species stock center, Bowling Green, OH. These sequences have been submitted to the GenBank database and have accession numbers L26039 for D. peninsularis Adh-Ψ and L26040 for D. mercatorum Adh-Ψ.

Isolation of D. hydei clone, SC-1, from a D. hydei genomic library in the vector EMBL-4 was accomplished for the D. hydei Adh clone, RS-7, and RNase protection, primer extension, and RNA sequencing were conducted as described elsewhere (Menotti-Raymond et al. 1991).
The nucleotide sequences of Adh pseudogenes from seven species, nine corresponding functional Adh genes, and the Adh gene of Scaptomyza albouvittata (used as an outgroup representative) were used to construct a gene tree by the following procedure. Introns of these genes were removed before the gene tree was constructed, because we intend to use the evolutionary tree derived from the intronless genes to evaluate intron evolution. In addition, introns have different rates of evolution compared with exons, and their alignment requires a number of insertions/deletions, some of which make comparisons arbitrary. The remaining nucleotides were aligned by the multiple alignment method of the program CLUSTAL V (Higgins et al. 1992). The dendrogram was produced by using Kimura’s two-parameter model (Kimura 1981) to estimate the number of nucleotide substitutions per site \((K)\) for all pairwise comparisons and by applying the neighbor-joining method (Saitou and Nei 1987) to construct the tree. Bootstrapping was accomplished by resampling the nucleotides, calculating \(K\), constructing the tree, and counting the number of times groups of sequences occurred together in 1,000 trials.

Codon biases in functional genes and pseudogenes were calculated by first aligning the pseudogenes to recreate the original reading frame and then counting the number of each type of base in the third position of each codon. The distribution of third-position nucleotides for the three genes (two in D. mulleri) was compared for each species by calculating the G-statistic, which is distributed as a \(\chi^2\).

Functional genes and pseudogenes of Adh were compared with respect to differences in the two introns by the following procedure. A phylogeny of the Adh genes was constructed for species of the mulleri and hydei subgroups from known sequences of homologous functional and homologous pseudo-Adh genes. Only the exons were used to construct the phylogeny by using both parsimony and maximum-likelihood methods (Felsenstein 1989). The resulting phylogeny for genes was identical to that estimated by Yum et al. (1991) with the addition of genes for D. buzzatii and D. mulleri. This phylogeny was used as a basis with which to determine the types and number of differences for the introns of both functional genes and pseudogenes. First, intron sequences were aligned by using the phylogeny to dictate the order of sequence pairs in the alignment procedure (i.e., the closest pairs in the phylogeny were aligned first, then these aligned sequences were used in the next step of the alignment procedure with the next closest sequence in the phylogeny, etc.). The multiple alignment sequence algorithm of the program CLUSTAL V (Higgins et al. 1992) was used with fixed and floating gap penalties of 10 and with weighted transitions. The aligned sequences of introns were then used to estimate the number of transitions, transversions, insertions, and deletions along each branch of the phylogeny, under the assumption that changes were parsimonious. This was accomplished by using the program DNAPARS of PHYLIP 3.4 (Felsenstein 1989) to reconstruct the changes along each branch of the user-defined tree. In those cases where character states were equally likely along different branches, a random assignment was made (except for deletions, which were considered less likely than substitutions). If there were contiguous deletions or insertions in the sequence along a single branch, then those events were considered to be singular (i.e., deletions at sites 4, 5, and 6 from node \(x\) to node \(x+1\) represented one deletion).

Synonymous and nonsynonymous nucleotide substitution rates were calculated by first aligning the pseudogenes to recreate the original reading frame. These aligned sequences were used to compare hypothetical codon positions for the pseudogenes and actual codon positions of Adh-2 genes (Adh in D. mulleri) by the method of Li et al. (1985). This analysis allows for an evaluation of paralogous sequences that have experienced the same amount of divergence time as either functional or nonfunctional genes.

**Results**

**Origin of the Adh Pseudogene**

In order to determine the point or points during evolution at which an ancestral Adh gene became inactivated to become a pseudogene, we have used nucleotide-sequence data from orthologous Adh pseudogenes from seven species representing each of four repleta subgroups. Sequences reported elsewhere include Drosophila hydei (hydei subgroup; Menotti-Raymond et al. 1991), D. mulleri (eremophila complex of the mulleri subgroup; Yum et al. 1991), D. mojavensis (mulleri complex of the mulleri subgroup; Atkinson et al. 1988), D. mulleri (mulleri complex of the mulleri subgroup; Fischer and Maniatis 1985) and D. buzzatii (mulleri complex of the mulleri subgroup; Shafer 1992).

There are two additional subgroups in the repleta group, mercatorum and repleta, for which Adh-\(\Psi\) sequences have not been reported. Consequently, we determined the nucleotide sequence of Adh-\(\Psi\) from D. mercatorum (mercatorum subgroup) and D. peninsularis (repleta subgroup). Drosophila peninsularis has a complete Adh-\(\Psi\), but the pseudogene of D. mercatorum contains two deletions in the Adh region. One of these is 36 bp long and is near the 5' end of exon 2. The other is approximately 180 bp and extends from exon 2 through the second intron to the beginning of the third exon. An amplified pseudogene fragment of D. paraanaensis, another species in the mercatorum subgroup, is also smaller.
than the amplified pseudogene fragments from other *repleta* group species and identical in size to the fragment obtained using *D. mercatorum* DNA (data not shown). It is thus likely that these deletions occurred in the lineage that formed the *mercatorum* subgroup.

We have aligned the 5' ends of seven *Adh*-Ψ sequences and compared them with an *Adh* gene known to be functional (data not shown) and found that each *Adh*-Ψ has an alteration very early in and disruptive to the *Adh* reading frame. *Drosophila pensularis* has an ACG instead of the translation initiation codon. *Drosophila pensularis*, *D. mulleri*, *D. mojavensis*, *D. hydei*, and *D. buzzattii* all have a single-nucleotide deletion in codon two. *Drosophila mettleri* has an alteration interpreted as a two base pair insertion at codon two. *Drosophila mercatorum* has multiple changes which involve codon 2. Therefore, if the definition of a pseudogene as a gene that is homologous to a known gene but is incapable of being correctly translated is used, then each of these genes is a pseudogene.

An additional alteration common to these species is a deletion of 600–1,000 bp 5' to the ancestral translation start codon. This deletion extends from a point 3' of the distal promoter, removes most of the large intron present in the untranslated region of an *Adh* primary transcript initiated from the distal promoter of the *Adh* gene in non-*repleta* species, and extends 3' to the proximal promoter region. This deletion would render each gene incapable of being translated during the larval stage. The presence of this deletion in genomic DNA can be evaluated by measuring the size of a PCR-amplified fragment that spans the deletion. Genomic DNA was amplified using a 5' primer whose sequence is located in a conserved region of the upstream adult enhancer (Ayer and Benyajati 1990; Sullivan et al. 1990) and a 3' primer whose sequence is located in exon 1. PCR amplification was performed with genomic DNAs from species outside and inside the *repleta* group as templates to identify species that have this deletion upstream of their *Adh* genes (fig. 2). The amplified DNA fragment from species with the deletion is expected to be approximately 600 bp smaller than that from species without the deletion. The amplified DNA fragment from *D. hydei*, which is known to have an *Adh* pseudogene, was ~0.5 kb. Amplified DNAs of about 1.1 kb were obtained from two species outside the *repleta* radiation, *D. virilis* and *D. robusta*, and from three species of the *repleta* section, *D. canalineae*, *D. Carmagoi*, and *D. gaucha*, which are not included within the *repleta* group. This suggests that each of these species has an *Adh* gene of ancestral structure, i.e., a coding gene with dual promoters. Amplified DNA fragments of ~0.5 kb were obtained from previously unanalyzed species of the *repleta* group, *D. fulvimaculata*, *D. pensularis*, *D. repleta*, *D. leonis*, and *D. mettleri*. Smaller but similarly sized products of about 300 bp resulted from amplification of DNA from *D. mercatorum* and *D. paraeneaen*.

These species are both in the *mercatorum* subgroup of the *repleta* group, and they apparently have a deletion larger than that of the other *repleta* group species or have multiple deletions. Therefore, the 5' deletion is confined to species within the *repleta* group. In all *repleta* group species that have been analyzed at the nucleotide sequence level, the *Adh*-Ψs have both the deletion and the alteration at codon 2.

Figure 3 shows the phylogenetic relationship of the *Adh* functional genes and pseudogenes, on the basis of nucleotide-sequence comparisons of the seven pseudogenes, nine corresponding coding genes, and the *Adh* gene of *Scaptomyza albivittata* (used as an outgroup representative). The result of this analysis shows that the pseudogenes form one group (95.5%) and the coding genes form another group (95.7%).
hypothesis that the Adh-Ψ's have a monophyletic origin prior to divergence of the repleta group species. We assume the inactivation of an Adh gene to become a pseudogene must have occurred after the initial Adh gene duplication since Drosophila without a functional Adh gene are not likely to survive. There have not, as yet, been extensive surveys of the subgenus Drosophila at the DNA level to determine the time of origin of the first Adh duplication. However, extensive surveys at the protein level have failed to identify species outside the repleta group that have isoforms of ADH (alcohol dehydrogenase 1.1.1.1) (Oakeshott et al. 1982; Batterham et al. 1984; Schafer 1992; D. T. Sullivan and P. Batterham, unpublished observations). Our data are all consistent with Adh-Ψ being at least as old as the repleta group, and conceivably it could be older, if species outside the repleta group had their origin after the initial Adh duplication. They do not have the 5′ deletion but could have an alteration that precludes translation, such as the changes at codon 2. Consequently, the latest time of origin for Adh-Ψ must be very early in the lineage leading to the repleta group.

Molecular Structure and Expression of the Adh Pseudogene

The initial Adh-Ψ examined was that from D. mulleri. It was reported that transcripts of Adh-Ψ were present in the adult stage. RNase protection assays indicated that the Adh-Ψ transcription start site was a short distance upstream from the ancestral translation start, at a position unrelated to any of the sequences homologous to those usually involved in Adh transcription initiation (Fischer and Maniatis 1985). Subsequently, using primer extensions in combination with RNase protection assays of the D. hydei Adh-Ψ transcript, we reported that this putative transcription start site was the 3′ boundary of an intron and that Adh-Ψ mRNA contained an additional 250 nucleotides derived from an exon located at an undetermined position upstream (Menotti-Raymond et al. 1991). In order to fully evaluate the possibility that Adh-Ψ has a function, it is necessary to fully characterize its molecular biology, to identify the transcription start position, and to characterize the 5′-most exon.

First, we determined the developmental history of Adh-Ψ transcripts compared with those of Adh-1 and Adh-2 by using RNase protection assays. These results are shown in figure 4. A 101-nucleotide fragment of an Adh-Ψ-specific in vitro synthesized transcript is protected by an RNA that is first barely detectable in very late third-instar larvae. This RNA increases in abundance early during pupation and shows an additional increase in relative abundance in adults. Of particular note is that this developmental profile is not at all similar to any known Adh transcript-expression pattern. Fragments of 83 and 62 nucleotides are also visible in figure 4. These represent fragments from both Adh-1 and Adh-2 mRNAs which are in part protected by Adh-Ψ transcripts. They show a pattern of accumulation typical of Adh transcripts, i.e., a rise during larval life, a decline during pupariation, and a second rise in the adult stage. In other analyses (S. W. Curtiss and M. Menotti-Raymond, unpublished data) we have used probes able to discriminate Adh-1 and Adh-2 transcripts in D. hydei are abundant in the first 4 d of larval life and decline to barely detectable levels thereafter, similar to the previously reported Adh-1 patterns in the development of D. mulleri (Fischer and Maniatis 1985) and D. mojavensis (Bayer et al. 1992).

Using the pattern of developmental accumulation of Adh-Ψ transcript as a guide, we examined the genomic region upstream from Adh-Ψ to identify the location of the transcripts that might contain the transcription start and 5′ exon of Adh-Ψ. We isolated from the D. hydei
Fig. 4.—RNase protection assays using a transcript prepared from a clone of Adh-Ψ to protect 20 µg total RNA prepared from Drosophila hydei of different ages in each lane. Lanes L, Larval stage. Lane 1, 64 h post oviposition. Lane 2, 88 h. Lane 3, 4.5 d. Lane 4, 5.5 d. Lane 5, 6.5 d. Lane 6, 7.5 d. Lanes P, Pupal stage. Lane 7, 1 d post pupation. Lane 8, 3 d. Lane 9, 4 d. Lane 10, 6 d. Lane 11, Adults (ad). Sizes were determined from an M13 sequence run on same gel (not shown).

Using reverse transcriptase for primer extension from the Adh-Ψ transcript, we generated a probe that was used to probe Southern blots of the two phages depicted in figure 5. Two fragments showed hybridization—the Xho-HindIII fragment from position 18–19 kb according to the map in figure 5, which contains the body of Adh-Ψ, and the HindIII-KpnI fragment, located at position 6 kb (data not shown). RNase protection assays (fig. 6) using a transcript synthesized from this fragment result in protection by an RNA that is present in larval, pupal, and adult stages at the expected relative abundances and is the size (252 nucleotides) expected for the exon upstream from Adh-Ψ. This suggested that the exon 5′ of Adh-Ψ is encoded in the HindIII-KpnI fragment at about 6 kb according to the map shown in figure 5. The nucleotide sequence of this fragment was determined and aligned with the sequence determined by RNA sequencing of the 5′end of the Adh-Ψ transcript (fig. 7). The two sequences are highly similar. The few discrepancies derive either from a very light C nucleotide lane in the RNA sequencing gel, which corresponds to the G nucleotide lane of this figure (see fig. 6 of Menotti-Raymond et al. 1991) or from mistakes in reading the number of identical bases at a position, which is not unexpected at the top of the gel (toward the left in fig. 7). The sequences become different at a point in the DNA sequence which is a good consensus for a 5′ splice site. The nucleotide sequence in the RNA 3′ of this point is found beginning at position 1736 of the D. hydei genomic sequence just upstream of the body of the Adh pseudogene. This position is immediately preceded by a 3′ splice site (Menotti-Raymond et al. 1991). Taken together, this body of evidence demonstrates that the transcription start site for the Adh-Ψ transcript is within the HindIII-KpnI fragment at 6 kb of figure 5. This
means that the intron spliced from the 5' region of the primary Adh-Y transcript is approximately 11.5 kb. The two introns within the Adh coding region are also spliced from Adh-Y in the usual manner of an Adh transcript (data not shown). We have thoroughly mapped the Adh-Y transcript only from D. hydei, but a 3' splice site is conserved in each Adh-Y sequence 13-19 bp upstream of the ancestral translation start site, and we know this splice site is used in D. mulleri and D. mojavensis (Fischer and Maniatis 1985; S. W. Curtiss and M. Menotti-Raymond, unpublished observations). This suggests that this 5' intron is present in all species having an Adh-Y.

Pseudogene Evolution

Figure 8 shows a comparison of three pairs of sequences, D. mojavensis versus D. mulleri, D. mojavensis versus D. buzzatii, and D. mojavensis versus D. hydei. These three species have close, moderate, and distant relationship to D. mojavensis, respectively. Accordingly, the increase in sequence divergence observed in the three comparisons reflects the longer divergence times for each species pair. Overall, the pseudogene is diverging only slightly faster than Adh-2 and is diverging much more slowly when compared with other regions, especially the intergenic region between the pseudogene and Adh-2. In addition, it is apparent from these similarity profiles that the pattern of divergence in exonic regions of both the pseudogene and Adh-2 are similar; that is, the exons are diverging more slowly than their neighboring introns. Neither the slow divergence in the pseudogene region relative to the intergenic region nor the similarity in the pattern of divergence for pseudogenes and functional genes is expected in comparisons involving a pseudogene.

We have made additional, more detailed comparisons of pseudogene and functional gene evolution. First, we compared coding genes and pseudogenes with respect to the relative rates of substitution at synonymous and nonsynonymous sites, after alignment and restoration of the reading frames to the pseudogenes. Table 1 lists the estimates for synonymous (Ks) and nonsynonymous (Ka) substitutions per nucleotide for functional genes and for hypothetical positions in the pseudogenes. Three types of comparisons were made. The first compares the synonymous rate with the nonsynonymous rate within each type of gene and for each species pair. In each case they are significantly different (P<0.05), indicating that synonymous sites are evolving faster than nonsynonymous sites in both functional genes and pseudogenes. This is the expected result when functional genes are compared. However, in comparisons involving pseudogenes, it is expected that divergence rates at synonymous and nonsynonymous sites of codons should be equivalent, if, as in this case, the species diverged after the inactivation event that created the pseudogene. A second comparison was between the synonymous-site rates of functional genes and pseudogenes. In these cases most species pairs (except D. mojavensis versus D. mulleri and D. mojavensis versus D. mettleri) show significantly (P<0.05) higher synonymous rates in the pseudogenes. This increase in rate of divergence at synonymous sites in the pseudogenes might be expected, if there has been a release from constraints of codon bias in the pseudogene. The third comparison was between nonsynonymous-site rates of the functional genes and pseudogenes. Even though all of the pseudogene functional gene-pair comparisons indicate a higher nonsynonymous rate in the pseudogene as compared with the functional gene, none is significantly different (P>0.05).
These results are not what would be expected in comparisons involving functional genes with pseudogenes. Once a gene is inactivated to become a pseudogene, comparison between pseudogene pairs and coding gene pairs of the same species should show a higher rate of divergence at nonsynonymous sites in the pseudogenes, unless the pseudogene is of very recent origin. Furthermore, there should be no difference in the rates of divergence at synonymous and nonsynonymous coding positions of the pseudogene, because coding is, by definition, not a property of pseudogenes.

To further evaluate the notion that there are no selective constraints on the pseudogenes, as compared with their paralogous functional genes, we have compared the frequency of codon usage or third-position bias in the two types of genes. Analysis of codon use by calculating scaled $\chi^2$ (Shields et al. 1988) shows that pseudogenes and functional genes of *D. mettleri*, *D. buzzatii*, and *D. mojavensis* have about the same codon-usage frequency. Only in the case of *D. hydei* is there a loss of codon bias in the pseudogene. This conclusion is reinforced by the data presented in table 2, which lists the number and percentage for each base at the third position for each species and type of gene. Again only *D. hydei* has a significant (G-statistic=16.8; degrees of freedom [df]=6; $P<0.01$) difference in the distribution of bases at the third position, with the pseudogene having a more even distribution than the functional genes. The *D. mulleri* comparison is not significant (G-statistic=8.2; df=6; $P>0.05$) but shows the same trend observed for the pseudogene of *D. hydei* (i.e., loss of codon bias in the pseudogene). This trend may indicate that selective forces that maintain codon bias have been somewhat relaxed.
We have compared the pseudogenes with the functional gene, Adh-2, with respect to the evolution of nucleotide sequences within the introns. After alignment of the introns of the functional genes and pseudogenes, intron 1 and intron 2 were 60 and 72 nucleotides in length, respectively. The number of changes, either base changes or deletions-insertions in all pairwise comparisons, are shown in table 3. Two-way G-statistics on the number of substitutions and insertions-deletions in intron 1 and intron 2 were calculated for each comparison and for each category of gene (functional or pseudo). None of the G-statistics was significant (P>0.05), thus the numbers for the two introns were pooled. Many of the comparisons of intron differences in functional genes versus pseudogenes are not independent because they contain the differences in early shared lineages. We present all of the possible comparisons to reveal any “recent” differences that might be present in the later lineages that are independent. In any given comparison, the time since divergence is the same; that is, the amount of time since D. mojavensis Adh-Ψ and D. mulleri Adh-Ψ diverged is the same as the amount of time since D. mojavensis Adh-2 and D. mulleri Adh-2 diverged, because the speciation event is the common departure in both cases. The data presented above argue for inactivation of the gene destined to become the pseudogene well before all speciation events that are relevant to these comparisons. We are thus comparing introns for genes under different selective constraints (i.e., functional genes vs. pseudogenes) for most if not all of their divergence time. Comparison of changes between functional genes of two different species involves three different cases. The first is the case where the functional genes duplicated after the species lineages diverged (i.e., after speciation.) An example of this case is the D. hydei genes versus the D. mojavensis genes. In this case the average number of changes for all four possible pairs of functional genes were calculated, to estimate amount of divergence in the functional gene introns. The second case is where the functional genes duplicated before speciation, such as Adh-1 and Adh-2 of D. mojavensis and D. mulleri. In this case only the two pairs of homologous functional genes of each species were compared. The third case involves the comparisons with the functional genes of D. mulleri. The average of the two possible comparisons was made in these cases.

Comparison of functional genes with pseudogenes for substitutions and insertions-deletions shows that, in general, the relative proportion of substitutions to insertions-deletions is about the same in the introns of the two types of genes (table 3). One pair (D. mulleri vs. D. hydei) was found to be significantly different. In this comparison there are about three times as many insertions-deletions separating the introns of their pseudogenes as there are separating introns of their functional genes.
In a separate analysis (Adh-1 and Adh-2 were not averaged in this analysis), maximum-likelihood estimates of branch lengths were added together, and the total length between functional gene pairs was tested for equality with total length for comparable pseudogene pairs. The test was done with a t-statistic based on the variance of the two lengths. Of 30 comparisons, 15 indicated no significant differences in the rate of sequence divergence in the introns of pseudogenes and functional genes, while 15 showed a significant increase (11) or decrease (4) in divergence of introns of pseudogenes relative to introns of functional genes. Most of the comparisons that showed significantly faster rates for pseudogene introns involved comparisons with D. hydei (i.e., D. mulleri with D. hydei [4], D. mojavensis with D. hydei [2], D. buzzatii with D. hydei [3], and D. mettleri with D. hydei [1]). Four cases showed the opposite (i.e., slower rates in the pseudogenes). These involved the introns of D. mojavensis Adh-2 with D. mulleri Adh-2 and D. mojavensis Adh-1 with D. mulleri Adh-1, and with both genes of D. buzzatii and D. mettleri. This relative conservation of introns in the pseudogenes as compared with the coding genes of D. mojavensis and D. mulleri is another unexpected feature of pseudogene evolution, first recognized by Atkinson et al. (1988).

We have also investigated intron and exon divergence by estimating \( K \) using Kimura's two-parameter method (transitions/transversions=2:1) for each exon and intron in the closely related pair D. mojavensis-D. mulleri and the more distantly related pair D. mojavensis-D. hydei. Coding gene versus pseudogene rates are shown with error bars (±1 standard error [SE]) in figure 9. These data show that, in general, the pseudogenes have diverged somewhat more than their comparable coding gene (Adh-2). As expected, the introns have diverged more than their neighboring exons; however, intron 2 sequences in the pseudogenes have either changed at the same rate as that of Adh-2 (D. mojavensis vs. D. hydei) or changed at a significantly slower rate than in Adh-2 (D. mojavensis vs. D. mulleri). This comparison
Table 3
Comparison of Functional Gene and Pseudogene Introns

<table>
<thead>
<tr>
<th>SPECIES PAIR</th>
<th>FUNCTIONAL GENES</th>
<th>PSEUDGENES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Substitutions</td>
<td>Indels*</td>
</tr>
<tr>
<td>D. mojavensis-D. mulleri</td>
<td>32.5</td>
<td>4.5</td>
</tr>
<tr>
<td>D. mojavensis-D. buzzatii</td>
<td>53.0</td>
<td>6.5</td>
</tr>
<tr>
<td>D. mojavensis-D. mettleri</td>
<td>64.0</td>
<td>15.0</td>
</tr>
<tr>
<td>D. mulleri-D. hydei</td>
<td>62.0</td>
<td>8.0</td>
</tr>
<tr>
<td>D. buzzatii-D. mettleri</td>
<td>36.5</td>
<td>5.0</td>
</tr>
<tr>
<td>D. mulleri-D. hydei</td>
<td>52.5</td>
<td>12.5</td>
</tr>
<tr>
<td>D. buzzatii-D. hydei</td>
<td>50.5</td>
<td>6.5</td>
</tr>
<tr>
<td>D. mettleri-D. hydei</td>
<td>53.0</td>
<td>13.5</td>
</tr>
<tr>
<td>D. buzzatii-D. hydei</td>
<td>51.0</td>
<td>6.5</td>
</tr>
<tr>
<td>D. mettleri-D. hydei</td>
<td>38.0</td>
<td>7.0</td>
</tr>
</tbody>
</table>

* Insertions and deletions, summed over both introns (132 nucleotides) for each of the 10 possible species pairs.

seems to indicate relative conservation in the intron 2 sequence and cannot be accounted for by a partial gene conversion from one of the two coding genes, because intron 2 of the two pseudogenes are more similar to one another than either is to their paralogous genes, Adh-1 or Adh-2.

Finally, we have attempted to address the question of possible Adh-Ψ function. We assume that, if a protein is encoded by Adh-Ψ and has a function and if this function is responsible for aspects of the evolution of Adh-Ψ not typical of a pseudogene, then we should find evidence of conservation of this function in the nucleotide sequences in the set of species that have diverged since the events that created Adh-Ψ. The potential for coding in the pseudogenes was investigated by determining whether any substantial open reading frame is common to the pseudogenes of all species. Figure 10 shows the location of termination codons for the three reading frames in a conceptual transcript from the pseudogene for each of six species. D. mercatorum Adh-Ψ was omitted from these comparisons because of deletions that remove substantial parts of exons 2 and 3. In each species in figure 10, translation is started at the first, second, or third base 3' of the presumed splice site in the 5' region. We chose this position because we have not experimentally determined the transcription start position, nor have we identified what might be a putative initiator methionine for all species. D. mercatorum Adh-Ψ was omitted from these comparisons because of deletions that remove substantial parts of exons 2 and 3. In each species in figure 10, translation is started at the first, second, or third base 3' of the presumed splice site in the 5' region. We chose this position because we have not experimentally determined the transcription start position, nor have we identified what might be a putative initiator methionine for all species. All species have a conserved 3' splice site 13-19 bases upstream of the original Adh start codon; and if a conserved open reading frame exists, then it would be evident in one of the frames beginning at this position.

Reading frames from the Adh-Ψ of D. mulleri, D. mojavensis, and D. mettleri (fig. 10) contain numerous termination codons, as a result of one or more frame-shifts. We have also evaluated the reading frames of these sequences with their introns intact and have not discovered any potential functional products. We have examined the opposite strand for open reading frames, and again we find no open reading frames that are common to this set of species. Long, uninterrupted reading frames are present in D. hydei and D. buzzatii sequences. Since the D. hydei transcription start region has been sequenced, we have also explored the coding capacity of the 5' exon. The HindIII-KpnI fragment shown in figure 5 is 933 nucleotides long. The 5' splice site for the 11.5 kb intron is 582 nucleotides to the right of the HindIII site. Primer-extension studies indicate that the 5' end of the transcript should be 252 nucleotides upstream from the splice site or about 330 bp to the right of the HindIII site. We inspected the sequence of this region and found a reasonable match to the Drosophila consensus transcription start sequence (Hultmark et al. 1986) beginning 331 nucleotides from the HindIII site. Downstream of nucleotide 331 there are two potential initiator methionine codons, one 59 nucleotides and the other 101 nucleotides 3' to the transcript 5' end. However, in frame with each of these is a stop codon. Thus, even in the case of one of the species whose Adh-Ψ has a long open reading frame, we find no evidence that the Adh-Ψ transcript can be productively translated.

Discussion

All of the evidence that is available indicates that Adh-Ψ has no function and, when any reasonable definition is used, is a pseudogene. Its origin as a pseudogene is very early in or just prior to the evolution of the repleta group in the subgenus Drosophila. In light of this relatively early origin, it is difficult to account for the level and types of nucleotide-sequence similarity and variation currently found in the Adh-Ψ of the species from throughout the repleta group. The following properties
lead us to favor the hypothesis that Adh-$\Psi$ in repleta group species does not have a function. There is no substantial open reading frame on either strand of the DNA that is common to the set of seven species. In the species that do retain a long open reading frame, Drosophila hydei and D. buzzatii, this reading frame would produce alcohol dehydrogenase (ADH) rather than another protein. Furthermore, in the case of D. hydei, we can find no methionine that might serve as an initiator codon, which is not separated from the open reading frame by a stop codon. We realize that this latter argument is extremely sensitive to possible sequencing errors. However, even if sequencing errors are assumed and a reading frame from a methionine is in frame with the downstream reading frame, the resultant protein would have $\sim$80 N-terminal amino acids followed by almost an entire ADH molecule. The function of this hypothetical protein seems uncertain. It seems likely that, since the ADH domain would be largely intact, at least some ADH catalytic activity might be retained. However, no additional ADH isoform in repleta species has ever been observed after substantial analysis (Batterham et al. 1982, 1984; Oakeshott et al. 1982; Schafer 1992; P. Batterham and D. T. Sullivan, unpublished observations). In addition, D. mercatorum has two deletions in the middle of its Adh-$\Psi$, and these would clearly abolish any ADH-related function. A suggestion that a function whose origin preceded the origin of the repleta group and now persists in some lineages, e.g., those of D. hydei and D. buzzatii which do retain a long open reading frame, but is lost from others necessitates a complex evolutionary argument that requires the improbable mutational inactivation at the same site (codon 2) independently in at least three separate lineages. Furthermore, this hypothesis does not address the basic result—that the pseudogene is evolving slowly—even in those species in which there is no long open reading frame.

There is at least one other transcript in the large 11.5-kb intron, and since this is on the strand opposite Adh-$\Psi$, the pseudogene is in the region 5’ to this gene.
It might be supposed that sequence conservation in the pseudogene is related to the maintenance of regulatory sites required for production of that transcript. However, the pattern of sequence conservation is not what might be expected for regulatory sites. We do not find blocks of sequence identity surrounded by regions of divergent stretches in the Adh-Ψs. Rather, there is a region ~1 kb long that is, on average, less divergent than expected over its entirety. It is possible that the transcript of the pseudogene has a function that is unrelated to its translation. We cannot rule this out. However, such a hypothesis cannot account for the pattern of sequence divergence of Adh-Ψ, specifically the slow rate of divergence at nonsynonymous codons.

Last, there appears to be heterogeneity with respect to the molecular evolution of Adh-Ψ within the set of seven species that we have analyzed here. Drosophila hydei, for example, has a very long, intact Adh reading frame. Yet, this is the species that has lost its Adh codon bias and seems to have the most rapidly evolving introns. On the other hand, D. mojavensis has numerous mutations that alter the Adh reading frame, yet its Adh-Ψ has retained much of the Adh codon bias. Drosophila mojavensis and D. mulleri also appear to have the slowest evolving introns, in the sense that the introns of D. mojavensis Adh-Ψ are more similar to those of D. mulleri Adh-Ψ, the closest two relatives among the species set examined here, than are the introns of Adh-2 of these two species.

Not many pseudogenes have been described in Drosophila, and there has only been one other study of the evolution of a Drosophila pseudogene (Jeffs and Ashburner 1991). This pseudogene is a processed pseudogene, which retains neither of its introns. These authors compared the evolution of the Adh pseudogene in two closely related species of the melanogaster subgroup, D. teissieri and D. yakuba. This pseudogene is not linked to its paralogue, and it is strongly argued, on the basis of shared molecular structure and identical chromosomal locations, that this Adh pseudogene arose before the divergence of these species. It is interesting that features of the evolution of this processed Adh pseudogene are reminiscent of features of the evolution of the Adh-Ψ in the repleta group. Except for a deletion that includes codon 1, there are no other frameshift deletions in the Adh reading frames. There is retention of codon bias, and, most notably, a comparison of the substitution rate at synonymous and nonsynonymous coding positions between the two pseudogenes reveals a $K_s/K_a$ ratio of 3.71. While not as pronounced as the $K_s/K_a$ ratios of the repleta Adh-Ψs, there is no obvious reason why this value should not be 1.0. Some of the peculiarities in the
Adh allows natural selection to provide a basis for the slow rate of evolution of the processed pseudogene may be due to its relatively recent origin. *Drosophila yakuba* and *D. teissieri* are more closely related than any of the *repleta* species that we have analyzed. The value of $K_3$ for the *Adh* coding genes of these species is about one-third of the value of $K_3$ between *D. mojavensis* and *D. mulleri*, the two closest relatives of the set analyzed here. On the other hand, we find it striking that in the only analyzed *Drosophila* pseudogenes, ones that have very different mechanisms of origin, similar puzzling features of molecular evolution are evident. More than likely, this reflects our lack of understanding of many aspects of the mechanisms that affect nucleotide-sequence origins, divergence, and conservation. Alternatively, it has been suggested that these *D. yakuba* and *D. teissieri Adh* sequences are not pseudogenes at all but are contributing to the structure of a gene that was created by the retrotransposition of *Adh* into an existing gene during their lineage (Long and Langley 1993). Shortly after pseudogenes were discovered it was believed they would serve as ideal models to study neutral evolution at the molecular level (Li et al. 1981). While this is likely true for some pseudogenes, it appears that this may not be a universal property of all pseudogenes.

For the reasons given above, we doubt that *Adh*-Ψ of the *repleta* group *Drosophila* has a function that would allow natural selection to provide a basis for the slow rate and peculiar pattern of nucleotide-sequence evolution. Consequently, we have sought other explanations. As a first approximation, it appears that the rate of *Adh*-Ψ evolution is only slightly faster than that of paralogous *Adh* coding genes. Clearly, the intergenic nucleotide sequences at the *Adh* locus are evolving substantially faster than the nucleotide sequences of the pseudogenes. In general, there are two hypotheses that might account for this slower-than-expected rate of evolution. First, the pseudogene might be located in a slowly diverging region of the genome. Second, the slow rate of evolution may be the result of a property related to the transcript or to the fact of transcription. There are several lines of data that can be used to support either alternative, and these alternatives are not mutually exclusive.

Studies using DNA-DNA reassociation kinetic analysis consistently have demonstrated that the *Drosophila* genome contains regions of single-copy DNA that evolve at different rates (Zweibel et al. 1982; Caccone et al. 1987, 1988, 1992; Caccone and Powell 1990; Werman et al. 1990; Powell et al. 1993). This conclusion has been derived from DNA-DNA reassociations by using DNA prepared from closely related species but also can be seen in intraspecific comparisons (Caccone et al. 1987). In intraspecific comparisons a substantial fraction of the genome fails to form duplexes at standard stringency of reaction. A specific example of slowly and rapidly evolving sequences in the genome of *D. melanogaster* at the cytogenetic region 68C has been identified (Martin and Meyerowitz 1986). Genes that encode glue proteins are located in this region. However, immediately adjacent to the glue-gene region is a region with a comparatively low rate of base substitutions in interspecific comparisons. The gradient between the slowly and rapidly evolving regions is steep, occurring over about 50 nucleotides. No function that might provide an explanation in terms of natural selection could be attributed to this slowly evolving region. It is conceivable that the *Adh*-Ψ is simply located in a region characterized by slow rate of evolution. However, we can find no specific properties through which the region described by Martin and Meyerowitz (1986) can be easily compared with the *Adh*-Ψ region.

Reassociation kinetic analysis using DNA fractionated into total single-copy cDNA (prepared using reverse transcriptase and poly A+ RNA) and intergenic DNA (DNA that remains after removal of the DNA that hybridizes with a 1,000-fold excess of poly A+ RNA) reveals that the cDNA fraction is part of the slowly evolving component. Because the cDNA fraction contains coding sequences, slower evolution resulting from the forces of natural selection seems reasonable. However, another factor that could conceivably contribute to slower evolution for the cDNA fraction is that the presence of transcripts or the act of transcription may either directly or indirectly affect the rate of sequence divergence of the transcribed regions. If transcripts can have an effect on DNA mutation and/or repair, then this might be relevant to the relatively slow divergence of a pseudogene that is transcribed. A number of direct or indirect mechanisms by which a transcript or transcription could influence the rate of nucleotide sequence divergence are plausible. At the present time we understand only the minimal aspects of the biochemistry of mutation and DNA repair and the ways it might be affected. We can presume that a transcript or a single-strand DNA reverse transcript copy could form a triple-helical structure or span a double-strand break, to result in increased fidelity of repair. Alternatively, the effect could be at the level of chromatin structure, where an open, transcriptionally active chromatin configuration might favor repair mechanisms of greater fidelity. While these remain highly speculative ideas, there are several observations that are consistent with a correlation between transcription and repair mechanisms or transcription and divergence of nucleotide sequences. First, it has recently been shown that the region of the mouse aprt gene, which is transcribed as the 3' untranslated region of the mRNA and which has no known function, evolves at a fourfold
lower rate as compared with the neighboring downstream untranscribed region (Turker et al. 1993). Second, by using mammalian cultured cells it has been shown that DNA repair, i.e., removal of thymidine dimers resulting from ultraviolet irradiation, is quantitatively much more extensive in transcribed regions (Mellon et al. 1986).

If a transcript or transcription is either directly or indirectly involved in biochemical events that affect nucleotide sequence divergence, this must occur in the germ line. In this regard, in the interspecific DNA-DNA reassociations cited above (Powell et al. 1993), cDNA sequences prepared from embryo poly A+ RNA were found to be less divergent than the comparable set of cDNA sequences prepared using poly A+ RNA from adults. If the presence of transcripts can slow the rate of nucleotide-sequence divergence through an effect on mutation or DNA repair, then the set of genes which encodes transcripts present during meiosis might show less sequence divergence than the set of genes whose transcripts are present only at other stages. Some of these may have been carried over into the embryonic RNA.

Both of these hypotheses regarding the relatively slow rate of evolution of Adh-Ψ are, of course, highly speculative. On the other hand, we know only the barest minimum about the biochemical events that are involved in DNA base substitutions in eukaryotic chromosomes in the germ line, the agents that may influence these events, or their eventual outcome. In any case, those principles that are commonly recognized as influencing the molecular evolution of genes appear inadequate to account for the molecular evolution of these Adh pseudogene sequences. This suggests that further investigation of both this system and others may reveal presently understood principles regarding the evolution of nucleotide sequences.

Acknowledgments

This research was supported by NIH grant GM-31857.

LITERATURE CITED


DANIEL L. HARTL, reviewing editor

Received October 5, 1993

Accepted January 21, 1994