Organization and Evolution of the Alcohol Dehydrogenase Gene in *Drosophila*[1]

Vivian H. Cohn[2] *and Gordon P. Moore*[3]
Division of Biological Sciences, University of Michigan, Ann Arbor

The alcohol dehydrogenase (*Adh*) gene was isolated from *Drosophila simulans* and *D. mauritiana,* and the DNA sequence of a 4.6-kb region, containing the structural gene and flanking sequence, was determined for each. These sequences were compared with the *Adh* region of *D. melanogaster* to characterize changes that occur in the *Drosophila* genome during evolution and to identify conserved sequences of functional importance. *Drosophila simulans* and *D. mauritiana Adh* are organized in a manner similar to that of *D. melanogaster Adh,* including the presence of two promoters for the single *Adh* gene. This study identified conserved flanking elements that, in conjunction with other studies, suggest regions that may be involved in the control of *Adh* expression. Inter- and intraspecies comparisons revealed differences in the kinds of sequence changes that have accumulated. Sequence divergence in and around the *Adh* gene was used to assess inter- and intraspecies evolutionary relationships. Finally, there appears to be an unrelated structural gene located directly 3′ of the *Adh* transcribed region.

Introduction

The alcohol dehydrogenase gene-enzyme system has been studied extensively by a variety of genetic, biochemical, and molecular means. Because it has been so well characterized, this system has been used as a model of eukaryotic gene-enzyme relationships. The alcohol dehydrogenase gene (*Adh*) is differentially expressed during development, and activity is concentrated in specific tissues (Ursprung et al. 1968, 1970). Some species-specific differences in expression have been described (Batterham et al. 1983; Dickinson et al. 1984; Fischer and Maniatis 1985).

The *Adh* gene of *Drosophila melanogaster* has been cloned, and the DNA sequences of the gene and flanking regions have been determined (Goldberg 1980; Benyajati et al. 1981; Kreitman 1983; Kreitman and Aguade 1986). We previously reported partial sequences for the *Adh* gene of *D. simulans* and *D. mauritiana* (Zwiebel et al. 1982; Cohn et al. 1984). Here, we extend this work by reporting the nucleotide sequence of 4.6 kb in each species, a sequence that includes the entire *Adh* gene and both 5′- and 3′-flanking DNA. This provides a comparison of 5′ regulatory sequences between *D. melanogaster* and these sibling species and expands the region for which sequences have been determined in these drosophilids.

The evolutionary relationships among these species have been examined by a

[1] Key words: alcohol dehydrogenase gene, DNA sequencing, *Drosophila* evolution, gene organization.

[2] Present address: Department of Neurology, University of Southern California, School of Medicine, MCH 142, 2025 Zonal Avenue, Los Angeles, California 90033.

[3] Present address: Integrated Genetics, 31 New York Avenue, Framingham, Massachusetts 01701.

Address for correspondence and reprints: Vivian H. Cohn, Department of Neurology, University of Southern California, School of Medicine, MCH 142, 2025 Zonal Avenue, Los Angeles, California 90033.

Mol. Biol. Evol. 5(2):154–166. 1988.

0737-4038/88/0502-0005$02.00

variety of methods (e.g., see Laird and McCarthy 1968; Zwiebel et al. 1982; Ashburner et al. 1984; Coyne 1984). All three species are closely related; indeed, crosses between *D. simulans* and *D. mauritiana* produce some fertile progeny. DNA sequence differences between these species are small and probably reflect primary changes. In the present study, differences are compared at the *Adh* locus, between alleles within *D. simulans* and *D. mauritiana,* and between these species.

Material and Methods

The *Adh* gene and flanking regions were isolated from genomic libraries of *Drosophila simulans* and *D. mauritiana* and subcloned into pBR325 (Zwiebel et al. 1982; Cohn et al. 1984). The subclones were designated pCAS (*D. simulans Adh*) and pCAM (*D. mauritiana Adh*). The 4.6-kb *Adh*-containing *Eco*RI fragments from each were separately cloned into M13, either as random fragments derived from sonication or as selected restriction fragments isolated from agarose gels. To generate random subclones, insert DNA from pCAS or pCAM was excised, self-ligated, and sonicated (Deininger 1983). Generated fragments were repaired using Klenow fragment (Boehringer-Mannheim) and then size-fractionated by electrophoresis in 1.5% agarose. Fragments of 250–600 bp were isolated and cloned into M13 mp10 DNA that had previously been digested with *Sma*I and treated with calf intestinal alkaline phosphatase (Boehringer-Mannheim). Transformation was carried out by standard techniques, using a modified strain of *E. coli* JM101 (TG1, provided by Dr. T. Gibson) which lacks the *Eco*K restriction system. Cells were made competent for the uptake of DNA by the method of Hanahan (1983).

DNA sequencing was carried out as described by Sanger et al. (1977). Samples were loaded onto 45-cm 6% acrylamide-buffer gradient gels (Biggin et al. 1983) and electrophoresed at 1,300 V, 30 mA for 3 h. Gels were fixed for 30 min in 10% acetic acid, dried onto 3MM chromatography paper, and autoradiographed with Kodak X-AR X-omat film at room temperature without intensifying screens.

On average, each nucleotide position was sequenced independently six times (fig. 1). Confirmation of sequence was obtained by sequencing overlapping clones, by repeated (or extended) sequencing of individual clones, and by sequencing both strands—with two exceptions: a region of 46 nucleotides and one of 160 nucleotides were sequenced on only one strand in pCAM. In these cases, the corresponding region of the other strand was sequenced several times using both overlapping and identical clones. Computer analysis of DNA sequence was carried out using programs available through Bionet. A+T-rich regions were defined as regions of 75% A+T in runs of eight or more nucleotides.

Results

A total of 4,607 nucleotides in the *Adh* region were determined in *Drosophila simulans,* and 4,581 were determined in *D. mauritiana.* The sequenced region extends from an *Eco*RI site ~1,300 nucleotides upstream from the adult promoter to an *Eco*RI site ~1,430 nucleotides downstream from the poly A addition site (fig. 1). The 26-bp difference in size between the sequenced region of the two species is the result of ⩾22 small deletions/insertions (fig. 2). The determined *D. melanogaster* sequence with which these are compared totals 3,977 nucleotides, from the 5′ *Eco*RI site to a position 801 nucleotides downstream from the poly A addition site. This sequence is derived from Adh^S alleles (Kreitman 1983; Kreitman and Aguade 1986).

There are two discrepancies between the chemical cleavage data and the sequence

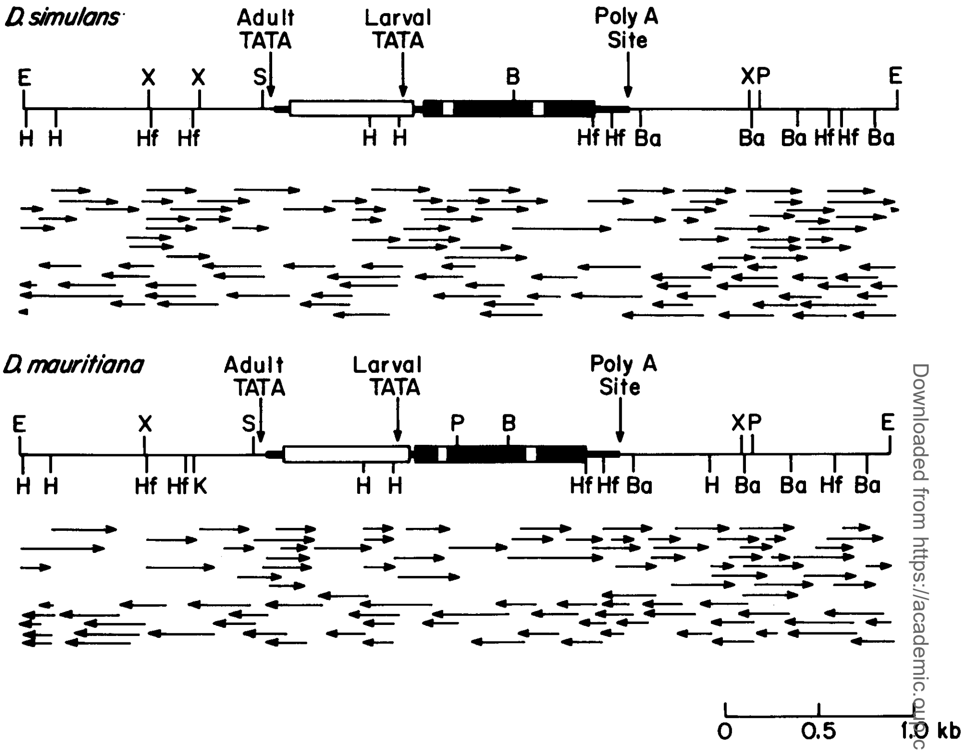

FIG. 1.—Strategy of DNA sequencing. A 4.6-kb *Eco*RI fragment of *Drosophila simulans* (*top*) and *D. mauritiana* (*bottom*) containing the *Adh* gene was cloned into M13 and sequenced by dideoxy chain termination. Arrows indicate the position, length, and orientation of sequenced clones. The restriction maps of the 4.6-kb *Eco*RI fragment are shown for each species in the 5′-to-3′ direction. The open-boxed regions are introns; the solid-boxed regions are exons; the thickened line indicates transcribed, nontranslated DNA.

determined by dideoxy chain termination. Positions 2225 (within intron 1 of pCAS and pCAM) and 2325 (in exon 2 of pCAS only) were determined to be T's by chemical cleavage or to be C's by dideoxy chain termination. It is likely that the chemical method is in error since it is more sensitive to base-chemistry and intrafragment interactions. These positions have been determined on both strands, several times, by dideoxy sequencing, and assignment of C is probably correct; all other *Drosophila* species and strains that have been sequenced have a C at these positions. In addition, there are five corrections to our previously published data. Position 1988 (position −117 in Cohn et al. 1984) is G rather than a deletion in both pCAS and pCAM, and position 1996 (−108) is an A in both species. Position 2148 (45) of pCAM is C; positions 2636–2647 (533–544) are ACCAAGGCCGCC in pCAS and pCAM; at position 2774 (672) both species have a T.

Organization of *Drosophila Adh*

Comparisons of the DNA sequence in the *Adh* region of *D. simulans* and *D. mauritiana* (present study) and in the *D. melanogaster* slow allele (Kreitman 1983; Kreitman and Aguade 1986) show them to be very similar and to share all major organizational features. In each species, the *Adh* gene is composed of three exons of 99, 405, and 267 bp separated by two introns averaging 66 and 68.5 bp, respectively.

```
         *         *         *         *         *         *         *         *         *         *         *         *
                 T                               A        ----                             G
GAATTCTTAAGCGTCTCCAAAGCTTCCAGAGAACTGGGCCATATATGACCCGATATCTGTATATATACATATATATATTTTAGCAAATATATTTTCCAATTAGGTGAAAGAACTTGTGTG    120

             ----------                                                                     T
CATACACACATATAGTTTTATATCAACAAACAGGTTTAAGTTTTATGCAAATTGAAAGCTTATTTCTTCCGCATGCTTATCTCTTTCCTTCTCATCATATGTATGTGATATTGCAGTAGC    240

                                       A      A                                                     G               T CC
CTCCTCCCACATCACATTTAACGCCCTATATTCACAATTTGCTCAAGAAAATATTTGAAACCAAATTGATTTCTAGTCAATTAGTTTTTAAGGCGCTGTGTGTCAGCTTAAAAAATGCAA    360

                                          C      T                     T
ATTAAGTGAGTAAAAATATAAACTTTTAATTCTTGCTAAACATATATACACATATATTTTTAATAAATAAACTACGAATATATAAAGACTACGAAGCTGGAAAACAAATTTAAAAATGAA    480
                     TTAA
                      ∇                                                       A                                    -
GCATTATAGTATTTTCGGTTTAATATATACATATGTATTAATCTAATGTGCGAGCATGGGTTAAATCTAGCTGCATTTGAAACCGCTACTCTGGCTCGGCACAAAGTGGGCTTGGTCGCA    600

                                                                                                      G        G
GTTGCGGACAAGTGAGATTGCTAATGAGCTGCTTTTAGGGGGCGTGTTGTGCTTGCTTTCCAACTTTTCTAGATTGATTCTACGCTGCCTCCAGCAGGCACCCCTCCCATCACCATCCCC    720

   -------------                  CT
ATCGCCATCCCCCATCACCATCCAGTCCCGCTGGTCCCCAGTCACAGTATTACACGTATGCAAATTAAGCCGAAGTTCAATTGCGACCGCAGCAACAACACGAGCTCGGATCTTTCTACA    840

                                                                -                  A              C AG
CTTCTCCTTGCTATGCTTGACATTCACAAGGTCAAAGCTCTTAGTATTCTGACTCGGGGTCCAAACACTGTAAAAAATTACTCTAGAAATAACGGTACACTTAATATGATTTTATTTTAA    960
                                                               TC
                  -----                                        ∇                             G
TCCATATGCCAAATATGTTATGTATGTTTATGTTATATTATTGTTAGGAAACCAATGTTTTTTTAAATCGGTTAAAAAGTTGCTACTAGAAAAACATAAACATTTTTGGAATATTTTTGA    1080
        T                                 T
        ∇ A            AT         A       ∇                                           A   T
ATATTTTCATCATTTTGTGTTTTTACGTTTTACTCATTTTTTTCTCAGTGCACTTTCTGGTGTTCCATTTTCTATTGGGCTCCTTACCCGCCATTTGTTTGCAAATCACTTGCTTGCGCA    1200

         A                            A        -                     A                                      A
TTTTTATTGCATTTTACATATTACAAATTATTTTTACGCCGCTGCTGCTGCCGCCGTCGACGTCGACTGCACTCGCCCCCACGACAGAACAGTATTTAAGGAGCTGCGGAGGTCCAAGTC    1320

                                                                                   C                        A
GCCGATTATTGTCTCAGTGCAGTTGTCAGTTGCAGTTCAGCAGACCGGCTAACGAGTACTTGCATCTCTTCAAATTTACTTAATTGATCAAGTAAGTAGCAAAAGGGCGACCAATTGAAG    1440
                                                                                  CAT                        TC
                                                            G                      ∇    ---                 - ∇       T
GAAATTCTTGTTTAATTTAATTTATTATGCAAGTGCGGAAATAAAATGACAGTATTAAACAGAAAATAAATTTTGCTAAAACTCATAAAATCAAATTTATTCAATCAAAATTAATTCAAA    1560
                                                            A
G      G G      A                         A              C C∇              A
CTGTCACAAATAGTGTGAACTCAATTAATTGGCATCTAATTAGAATTTGGAGGCCTGTGTCGAATATTCGTCTTGGAAAATCACCTGTTAGTTAACTTCTAAAAATAGGAATTTTAACAT    1680
                                                                                        CAGGCAT
        T               A                                                                  ∇
AACCCGCCCCTGTTAATCGGCGCCGTGCCTTCGTTAGCTATCTCAAAAGCGAGCGCGTGCAGACGAGCAGTAATTTTCCAAGCATCAGGCATAGTTGGGCATAAATTATAAACATACAAA    1800

                                                          C      G                      C
CCGAATACTAATATAGAAAAAGCTTTGCCGGTACAAAATCCCAAACAAAAACAAACCGTGTGTGCCGAAAAATAAAAATAAACCATAAACTAGGCAGCGCTGCCGTCGCCGGCTGAGCAG    1920

  T
CCCGCGTACATAGCCGAGATCGCGTAACGGTAGATAATGAAAAGCTCTACGTAACCGAAGCTTCTGCGGTACGGATCTTCCTATAAATACGGGGCCGACACGAACTGGAAACCAACAACT    2040

                                           A                                                               C           T
AACGGAGCCCTCTTCCCATTGAAACAGATCGAAAGAGCCTGCTTAAGCAAAAAAGAAGTCACCATGGCGTTTACTTTGACCAACAAGAACGTGATTTTCGTTGCCGGTCTGGGAGGCATC    2160

                        G                                                   --  G            A       T               G
GGTCTGGACACCAGCAAGGAGCTGCTCAAGCGCGACCTGAAGGTAACTATGCGATGCCCACAGGCTCCATGGAGGGGGATTGAGGTTAATCTCGTGTATTCAATCCTAGAACCTGGTAAT    2280

                     T                  C
CCTCGACCGCATTGAGAACCCGGCTGCCATTGCCGAGCTGAAGGCAATCAATCCAAAGGTGACCGTCACCTTCTACCCCTATGATGTGACCGTGCCCATTGCCGAGACCACCAAGCTGCT    2400

GAAGACCATCTTCGCCAAGCTGAAGACCGTCGATGTCCTGATCAACGGAGCTGGTATCCTGGACGATCACCAGATCGAGCGCACCATTGCCGTCAACTACACTGGCCTGGTCAACACCAC    2520

GACGGCCATTCTGGACTTCTGGGACAAGCGCAAGGGTGGTCCCGGTGGTATCATCTGCAACATTGGATCCGTCACTGGTTTCAATGCCATCTACCAGGTGCCCGTCTACTCCGGCACCAA    2640
                                                                       G
                                                               --      ∇ C
GGCCGCCGTGGTCAACTTCACCAGCTCCCTGGCGGTAAGTTCATCGAAGGAAACGCAAAGTTTTCAAAAAATAAAAAACTAATTTGTTTTATAACACCTTTAGAAACTGGCCCCCATTAC    2760

                                                                  A
CGGCGTGACCGCTTACACCGTGAACCCCGGCATCACCCGCACCACCCTGGTGCACAAGTTCAACTCCTGGCTGGATGTTGAGCCCCAGGTTGCCGAGAAGCTCCTGGCTCATCCCACCCA    2880

      A                                                     C                  A                G
GCCCTCGTTGGCCTGCGCCGAGAACTTCGTCAAGGCTATCGAGCTGAACCAGAACGGAGCGATCTGGAAACTGGACTTGGGCACCCTGGAGGCCATCCAGTGGACCAAGCACTGGGACTC    3000

CGGCATCTAAGAAGTGATACTCCCAAGAAACACATAACATTAGTTCATAGGGTTCTGCGAACCACAAGATATTCACGCAAGGCAATAAGGCTGATTCGAAGCACACTCACATTCTTCCCC    3120

                                                                                                     A
TAATACGATAATAAAACTCTCCATGAAAAATATGGAAAAATATATGAAAATTGAGAAATCCAAAAAACTGATAAACGCTCTACTTAATTAAAATAGATAAATGGGAGCGGCTGGAATGGC    3240

                                                                                  G
GGAACATGGCCAAGTTCCTCCGCCAATCAGTCGTAAAACAGAAGTCGTGGAAAGCGGATAGAAAGAATGTTCGATTTGACGAGCAAGCATGTCTGCTATGTGGCGGATTGCGGAGGAATT    3360

                                                             G                         G
GCACTGGAGACCAGCAAGGTTCTCATGACCAAGAACATAGCGGTGAGTGAGCGGGAAGCTGACGGCTTTGGGGTTCGGTTTCTGTCCAGATCGAACTCAAAACTAGTCCAGCCAGTCGCT    3480
                                                                                   TACGCT
                                   A                                                ∇    A                       C
GTCGAAACTAATTAAGTTAATGAGTTTTTCATGTTTGTTTCGCGCTGAGCAACAATTAAGTTTATGTTTCAGTTCAGTTTTGATGAAGGACTTGGCACTTTCAATGAATACTTTAGAACA    3600

   T                                     --         C
AAAGCACAACTCATTCTAATAACTTGGTGGTCATATTTTTTTTAATGATAAGTTTAAACTTTTTATATTTCGATATTAAACCACCTATGAAGTCTATTTTAATCGCCAGATAAGCCATAT    3720

      T                                             G                                   --------
ATACTGCGTAAATATTTGTATTCTTTATCAGGAAATTCAGAGAAACGGGGAATTTACTATCTACTAAAAGCCAAACAATTTCTTACAGTTTTACTCTCTCCACTCTAGAAATTGGCCATT    3840

                                                  G  A     T                                       A
TTACAGAGCACGGAAAATCCCCAGGCCATCGCTCAGCTGCAGTCGATAAAACCGAGTACCCAAATATTTTTCTGGACCTACGACGTGACCATGGCAAGGGAAGAAATGAAGAAGTACTTC    3960

GATGAGGTGATGGTCCAAATGGACTACATCGATGTCCTGATCAATGGAGCTACGCTGTGCGATGAAAATAACATTGATGCCACCATCAATACAAATCTAACGGGAATGATGAACACTGTG    4080

        G                                                           G                                   C
GCCACGGTATTACCCTATATGGACAGAAAAATGGAGGATCTGGTGGGCTAATTGTGAACGTTACCTCCGTCATTGGATTGGACCCTTCGCCGGTTTTCTGCGCATATAGTGCATCCAAG    4200

                                           G           A A       A                -
TTCGGTGTGATTGGATTTACCAGAAGTCTAGCGGTGAATTGAATCCAATCTTATCGGGATAAATTCTTTAATTTTTGCTTTTTAGGACCCCCTTTACTATTCCCAAAACGGAGTCGCTGT    4320

                     T                     G                A                                G
GATGGCGGTTTGTTGTGGTCCCACAAGGGTCTTTGTGGACCGGAAACTGAAAGCGTTTTTGGAATACGGACAATCCTTTGCCGATCGCCTGCGTCGAGCTCCCTGCCAATCGACATCGGT    4440

GTGTGGTCAGAATATTGTCAATGCCATCGAGAGATCGGAGAATGGCCAGATATGGATTGCGGATAAGGGTGGACTGGAGTTGGTCAAATTGCATTGGTACTGGCACATGGCCGACCAGTT    4560

CGTGCACTATATGCAGAGCAATGATGAAGAGGATCAAGATTGAATTC
```

FIG. 2.—DNA sequence of the *Adh* gene and flanking regions. The DNA sequence of *Drosophila simulans Adh* is shown; differences found in *D. mauritiana* are indicated above the line. Deletions are represented by a dash; insertions are indicated by an inverted triangle.

Table 1
Positions of Structural Features in the *Adh* Genes of *Drosophila simulans* and *D. mauritiana*

	Nucleotide Position*	
Feature	pCAS	pCAM
Adult TATA box	1293	1266
Adult mRNA	1325–3181	1298–3160
Adult intron	1412–2067	1385–2049
Larval TATA box	2002	1984
Larval mRNA	2034–3181	2016–3160
Exon 1	2104–2202	2086–2184
Intron 1	2203–2269	2185–2249
Exon 2	2270–2674	2250–2654
Intron 2	2675–2743	2655–2722
Exon 3	2744–3010	2723–2989
Poly A recognition site	3130	3109
Poly A addition site	3181	3160

* Bases are numbered from the 5′ end of the *Eco*RI fragment.

The intron/exon borders contain the consensus splice signals described by Breathnach et al. (1978). Some 102 nucleotides upstream from the translation-initiation ATG codon is the sequence TATAAATA, identical in all three species, that corresponds to the TATA box commonly found upstream from eukaryotic genes (Goldberg 1979). The distance between this region and the ATG site is that expected on the basis of the length of the 5′ leader sequence of *Adh* mRNA in *D. melanogaster* larvae (Benyajati et al. 1980). A second TATA box is found ~811 nucleotides upstream from the ATG site. This sequence, TATTTAA, is again identical in all three species. In *D. melanogaster* this putative promoter has been shown to be necessary for transcription of *Adh* in the adult fly (Benyajati et al. 1983; J. Posakony, personal communication). An additional intron has been identified in the region between the "adult promoter" and the ATG site. The "larval promoter" is contained within this intron. Downstream from the gene in all three species there are 171 nucleotides of nontranslated, transcribed DNA, ending at AATCC (Benyajati and Dray 1984) (table 1). It should be noted that message size, transcription initiation, and termination positions were determined for *D. melanogaster* (Benyajati et al. 1981) and are identified in *D. simulans* and *D. mauritiana* by alignment of these homologous sequences.

Distribution of Differences in the *Adh* Region

To study the level of sequence constraint in various regions during evolution, we examined, for each section of the *Adh* locus, the amount and types of differences that exist between the species. In this analysis, all differences were weighed equally, and deletion/insertions were scored as single events irrespective of their length. Table 2 shows that the greatest sequence change occurs in the 5′-nontranscribed DNA and in introns 1 and 2; the "adult intron" is more conserved than the others in *D. melanogaster* relative to the sibling species (by a paired *t*-statistic test, $P < .04$). The 5′ leader sequences have diverged to approximately the same extent as coding DNA, i.e., approximately one-third as much as the introns. The 3′-nontranslated DNA has diverged significantly in *D. melanogaster* relative to *D. simulans* and *D. mauritiana* ($P < .005$ by a *Z*-statistic

Table 2
Distribution of DNA Sequence Differences in the Adh Region of Three Drosophilids

	% DIVERGENCE* (Species or Alleles Compared)				
REGION (No. of Sites Compared)	si^a/me^s	si^a/ma^a	ma^a/me^s	si^a/si^b	ma^a/ma^b
5′ Nontranscribed (1,407 [47])	7.9	3.1	8.8	2.1	0.0
Adult mRNA leader (124)	1.6	1.6	1.6	1.6	0.8
Larval mRNA leader (70)	2.9	1.4	1.4	1.4	0.0
Adult intron (668)	3.9	3.3	5.4	2.7	3.4
Intron 1 (67)	6.0	6.0	6.0	1.5	4.5
Intron 2 (70)	10.0	4.3	11.4	2.9	4.3
Exon 1 (99)	3.0	3.0	4.0	0.0	0.0
Exon 2 (405)	1.0	0.7	1.2	0.5	0.0
Exon 3 (267)	1.5	1.9	2.6	0.4	0.4
3′ Nontranslated (171)	5.3	0.0	5.3	. . .	. . .
3′ Nontranscribed (826 [1,451]†)	4.2	2.0†	4.5	. . .	. . .
Total introns (805)	4.6	3.6	6.0	2.6	3.6
Total exons (717)	1.4	1.4	2.1	0.4	0.1
Overall (4,104 [4,729]† [1,747]‡)	5.0	2.5†	5.8	1.5	1.8

NOTE.—si = *Drosophila simulans*; ma = *D. mauritiana*; and me = *D. melanogaster*. Superscript letters a, b, and s denote source of data, as follows: a, present study; b, Bodmer and Ashburner (1984); s, Kreitman (1983) or Kreitman and Aquade (1986).

* Scoring base changes and insertions/deletions.

† In the 3′-nontranscribed region, 1,451 nucleotides of *D. simulans* and *D. mauritiana* were compared. In the first 826 nucleotides of this region all three species were compared, and these two species were found to differ by 2.2%. In the overall measurement, the value of 2.5% divergence was the same whether compared over 4,104 or 4,729 nucleotide positions.

‡ Number of sites compared between alleles.

test), while the latter have not diverged at all relative to each other. The 3′-nontranslated region has diverged 3½ times more rapidly than the 5′ leader in *D. melanogaster* compared with each of the sibling species. The 3′-flanking (nontranscribed) DNA has diverged slightly less than 3′-transcribed, nontranslated DNA ($P < .05$) and approximately two-thirds as much as the 5′-nontranscribed region ($P < .05$). If the same comparisons are confined to single base substitutions only, the absolute number of differences decreases, but relative trends are maintained, though the 3′-nontranslated, 3′-nontranscribed, and adult intron regions appear to have diverged to an equal extent.

Sequence differences are not evenly distributed within each structural region (fig. 3). Changes are clustered, especially in the 5′-nontranscribed region, in the adult intron, and, to a lesser extent, in the 3′-nontranscribed region. In the adult intron in particular, different regions have diverged to different extents, an observation that suggests there may be sequences of functional importance here. Deletion studies (Goldberg et al. 1983; J. Posakony, personal communication) have been used to identify possible regulatory regions. Examination of the region suggested to contain larval regulatory elements (from *D. simulans* positions 1654–2002) reveals strikingly fewer base changes than those seen in the surrounding DNA (fig. 2). In interspecies comparisons, this region is 2.6% divergent, while the remainder of the adult intron differs at 12.8% of the nucleotides. The region suggested to control adult-specific expression (positions 902–1165) is much less conserved, but there is a perfectly conserved run of 39 nucleotides (from *D. simulans* position 1124; CTCAGTGCACTTTCTGGTGTT-

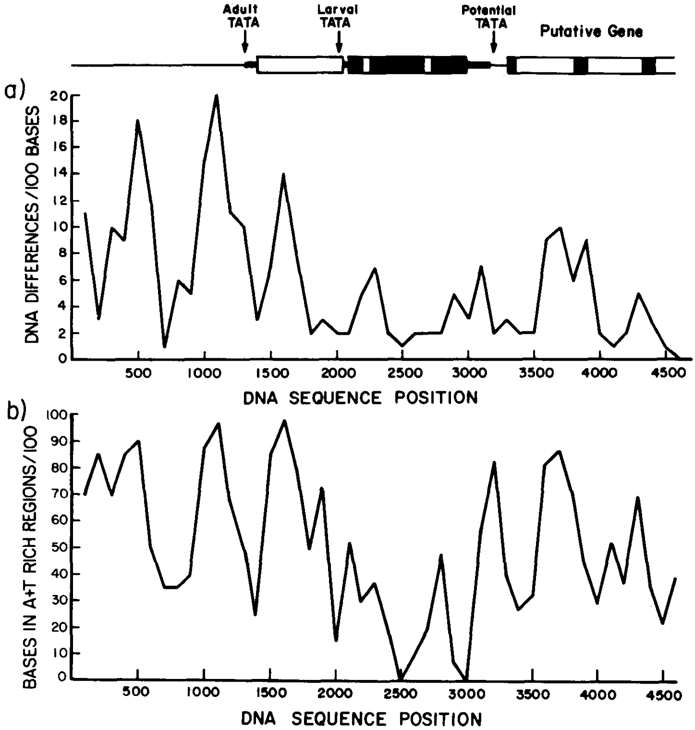

FIG. 3.—Distribution of DNA differences and A+T-rich regions in and around the *Adh* gene. The number of DNA differences/100 bases (a) and the number of bases in A+T-rich regions/100 nucleotides (b) are plotted vs. DNA sequence position. A+T-rich regions are regions with >75% A+T in runs of at least eight nucleotides. The approximate location of a putative gene 3′ to *Adh* is shown; the second exon of this gene may extend to position 4234, although the shortest size consistent with the data is indicated. Open boxes are introns, and solid boxes are exons.

CCATTTTCTATTGGGCTC) and one of 22 nucleotides (from position 986; GTTTATGTTATATTATTGTTAG). Between positions 1165 and 1228, where truncation also results in aberrant adult control, there is a 53-bp run with only two differing sites. Interestingly, an inverted repeat lies within this region at positions 1228–1239. It will require additional deletion studies, perhaps focusing specifically on these conserved regions, in conjunction with sequence comparisons of more distant species to define the control regions further.

Characterization of Divergence in the *Adh* Region

Table 3 lists changes in exons of *Adh* according to their positions in codons. There are twice as many synonymous differences as those leading to amino acid replacements. When the method of Perler et al. (1980) is used to correct for the relative number of silent and replacement opportunities in each codon, the ratio of silent to replacement changes increases by another 2.5-fold.

Coding-region differences are restricted to single nucleotide differences. Noncoding DNA, however, contains clustered differences and insertions/deletions as well as scattered single base differences. The ratio of insertions/deletions to base changes in noncoding DNA is 0.27, which is similar to the value of 0.28 reported by Cann and Wilson (1983) for many noncoding DNAs. Kreitman (1983) observed a ratio of 0.26 among

Table 3
Differences in the *Adh* Coding Regions of Three Drosophilids

	No. of Differences					Replacement/ Silent Substitutions	
	Codon Position						
Species Compared*	1	2	3	Replacement	Silent	Ratio	Corrected Ratio†
Drosophila simulans/D. melanogaster	3	0	8	2	9	0.22	0.07
D. simulans/D. mauritiana	4	0	7	4	7	0.57	0.23
D. mauritiana/D. melanogaster	7	0	9	6	10	0.60	0.20

* *D. simulans* and *D. mauritiana* sequences compared are those alleles reported here; *D. melanogaster* is allele Af-S (Kreitman 1983).

† Corrected, by the method of Perler et al. (1980), for the expected relative frequencies of random mutations causing replacement and silent substitutions.

D. melanogaster alleles, while Bodmer and Ashburner (1984) found a value of 0.22 when comparing the 5′-flanking region and introns of the *Adh* gene from several drosophilids.

Interspecies Variation in the *Adh* Region

As shown in table 2, the *Adh* regions of *D. simulans* and *D. mauritiana* are clearly more similar to each other than either is to *D. melanogaster.* This is most apparent in the 5′-nontranscribed region, intron 2, the 3′-nontranslated region, and the 3′-nontranscribed DNA (by a *t*-statistic test for two means, $P < .05$). Overall, *D. simulans* and *D. mauritiana* differ by 2.5%, *D. simulans* and *D. melanogaster* differ by 5.0%, and *D. mauritiana* and *D. melanogaster* differ by 5.8%. When only base substitutions are scored, these values are 2.0%, 3.9%, and 4.7%, respectively.

The rate of divergence of single-copy DNA in *Drosophila* was previously estimated to be ≥0.66% of bases/Myr (Zwiebel et al. 1982). Hunt et al. (1981) estimated 0.4% for Hawaiian drosophilids. This rate can be applied to the extent of divergence reported here to yield an estimate of the times of divergence of the species. Obviously, application of an overall divergence rate to a particular locus yields a very rough, and possibly misleading, estimate. Nevertheless, the sequenced region is large enough that such an estimate may have some validity. Application of this rate to the *Adh* data, in a region of 1,747 nucleotides common to the sequences, yields estimated divergence times of 4.5, 3.8, and 5.9 Myr for *D. simulans/D. melanogaster, D. simulans/D. mauritiana,* and *D. mauritiana/D. melanogaster,* respectively. Table 4 shows the percent differences among the *Adh* alleles in the region of overlapping sequence (1,747 nucleotides). With these values, a phylogeny was constructed by using the method of Fitch and Margoliash (1967) (fig. 4).

Polymorphic Variation in the *Adh* Region

A portion of the 4.6-kb region has been sequenced by others from separate strains of *D. simulans* and *D. mauritiana* (Bodmer and Ashburner 1984). Thus, polymorphism between these alleles can be assessed in the *Adh* region (table 2). Bodmer and Ashburner (1984) sequenced from position 1278, 47 nucleotides upstream from the adult transcription initation site, to position 3012, the translation-termination codon. There are very few polymorphic differences in coding DNA. Of the four differences between

Table 4
Divergence of *Drosophila Adh* Alleles*

	Me^s	Me^f	Si^a	Si^b	Ma^a	Ma^b
Me^s		1.2	3.0	2.9	3.9	3.3
Me^f	1.2		3.7	3.4	4.6	3.9
Si^a	3.0	3.6		1.6	2.5	2.5
Si^b	3.0	3.6	1.6		2.6	2.0
Ma^a	3.8	4.4	2.6	2.6		1.7
Ma^b	3.1	4.1	2.3	2.3	1.7	

NOTE.—The divergences (as a percentage of the nucleotide positions that differ) are shown in the upper right half of the table. In the lower left half of the table are the phyletic distances obtained by summing the distances on the branches of the tree in fig. 4 that connect two taxa. The disagreement between the two halves of the table, as represented by the % SD statistic (Fitch and Margoliash 1967), is a reasonably low 5.6 and would have been lower if the computations had been carried beyond the first decimal.

* Species and allele designations as in fig. 4 below.

sequenced alleles, the three changes in *D. simulans* are silent while the one in *D. mauritiana,* at nucleotide 2454, results in replacement of isoleucine by valine. Interestingly, the kinds of changes found in the polymorphic comparisons are somewhat different from those observed between species. Between alleles, differences are fewer but clustered; 40.4% of the polymorphic sites lie within three nucleotides of one another, while only 25% of the *D. simulans*/*D. mauritiana* differences in the same region fall in this category. Some 60% of the observed differences have accumulated in the first

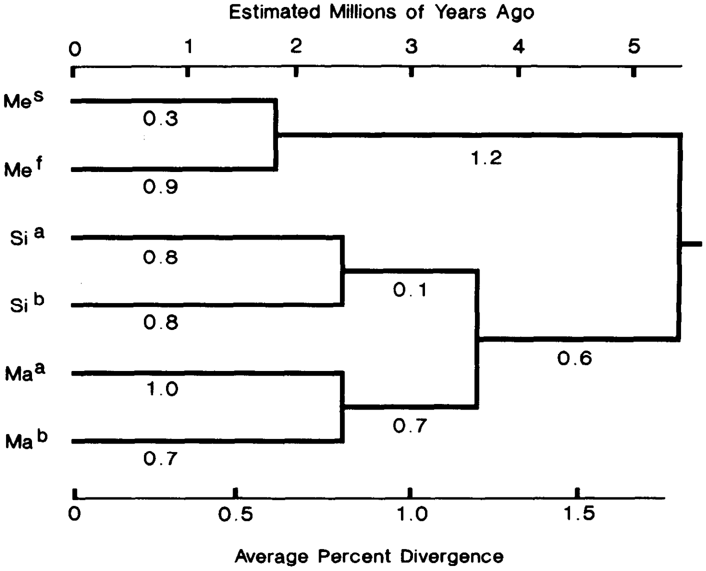

FIG. 4.—Phylogeny of *Drosophila Adh* genes. The Fitch and Margoliash (1967) procedure was used to contruct the phylogeny from the divergence values in the upper right half of table 4. Ancestral nodes are plotted, using the lower scale, at the average of the divergences in the weighted lines of descent from that node. The actual percent divergence estimated in an internodal interval is given adjacent to the line representing that interval. The taxa are as follows: Me = *D. melanogaster;* Si = *D. simulans;* and Ma = *D. mauritiana.* The alleles are either slow and fast (s and f, respectively) or a (present study) and b (Bodner and Ashburner 1984). The time scale (upper scale) is drawn assuming a rate of 3.3 × 10^{-9} nucleotide substitutions/site/year.

226 nucleotides of the adult intron. The length polymorphisms are very short, one to four nucleotides only, and most are single base insertions/deletions. The ratio of insertions/deletions to base changes in noncoding DNA is 0.13 for *D. simulans* and 0.29 for *D. mauritiana,* or 0.21 for both together. These ratios are similar to those observed for the interspecies noncoding DNA comparisons. The levels of intraspecies divergence observed are similar to that seen between *D. simulans* and *D. mauritiana* (1.6% or 1.7% vs. 2.5%). Thus the branching time is likely to be similar as well. This would militate against the possibility that differences in the type and distribution of nucleotide changes are due to divergence time alone.

Evidence for a Gene 3′ of *Adh*

The overall base composition of the *D. simulans* sequenced region is 29.7% A, 22.2% C, 20.6% G, and 27.5% T (*D. mauritiana* is 29.9% A, 22.0% C, 20.7% G, and 27.4% T). However, A+T residues are distributed unequally (fig. 3). As is the case in most genes, coding DNA is less A+T-rich than noncoding regions (Gojobori et al. 1983). Interestingly, the 3′-nontranscribed DNA appears to have the A+T distribution characteristic of a gene. Moreover, there are several small open reading frames, separated by apparent introns that contain the consensus splice signals. A+T content is highest in the putative introns and lowest in the putative exons. Further, the distribution of base changes and insertions/deletions correlates with the proposed location of this gene; that is, the exons have accumulated only nucleotide substitutions, while length changes have occurred in the introns. A possible promoter, TAATTAAAATA, is located 24 bases from the end of the *Adh* poly A addition site and is conserved in all three species. An open reading frame begins ~100 nucleotides further downstream at the methionine codon.

Discussion

Identification of Regulatory Elements

In the *mulleri* subgroup of *Drosophila* there are two *Adh* genes, expressed differentially during development (Oakeshott et al. 1982; Batterham et al. 1983; Fischer and Maniatis 1985). The single *Adh* gene of *D. melanogaster* is controlled by two promoters. The developmental specificity of *Adh* expression is reflected in the specificity of transcription initiation at the two promoters (Benyajati et al. 1983). Based on the phylogenetic distribution of these (or additional) organizational patterns plus genomic Southern blots (Zwiebel et al. 1982), *D. simulans* and *D. mauritiana* also appear to possess a single *Adh* gene with two promoters. This is not surprising, since *D. melanogaster, D. simulans,* and *D. mauritiana* are closely related, as judged on the basis of many molecular and nonmolecular characteristics (reviewed, e.g., in Ashburner et al. 1984).

Evolution of the *Drosophila Adh* Gene

The relatedness of total single-copy DNA of *D. melanogaster* to that of *D. simulans* or *D. mauritiana* was assessed elsewhere by reassociation and thermal denaturation (Zwiebel et al. 1982), yielding values of 1.9% and 2.3% divergence, respectively. Considering the observed divergence of the 4.6-kb *Eco*RI fragment (table 2), it appears that single-copy DNA is diverging at approximately the same rate as *Adh*-coding DNA but approximately one-third the rate of noncoding regions. The fact that total single-copy DNA seems as conserved as that coding for *Adh* suggests that much of the genome, while serving no coding function (reviewed, e.g., in Moore 1984), may play a sequence-dependent regulatory or structural role rather than being free to diverge randomly.

As described previously, interspecies *Adh* comparisons revealed nucleotide substitutions and length polymorphisms, the latter restricted to noncoding regions. The differences are fewer and more highly clustered in intraspecies comparisons, and the length polymorphisms are very short. The clearest difference between inter- and intraspecies comparisons is in the number of amino acid replacements. Kreitman (1983) compared 11 strains of *D. melanogaster* and observed 14 polymorphisms in *Adh*-coding DNA. All but one of these were silent, with the one replacement substitution accounting for the threonine-to-lysine difference that distinguishes *Adh*$^{\mathrm{F}}$ and *Adh*$^{\mathrm{S}}$ alleles. The high number of silent polymorphisms led Kreitman to suggest that virtually all amino acid replacements within this species have been selected against. The two *D. simulans* alleles compared in this study have a 3:0 ratio of silent:amino acid replacement changes, while a ratio of 0:1 is found for the *D. mauritiana* alleles. However, among the 14 codons that vary between species, nine are silent and five are replacement differences. (When all six alleles are considered, a ratio of 17 silent:7 replacements is observed). If it is true that no changes are neutral, this high fraction of replacements between species suggests that environmental constraints on the three species are different. It is clear from a phylogenetic comparison (fig. 4) that the lineages have evolved at unequal rates since their separation. In fact, at the *Adh* locus, *D. mauritiana* appears to have evolved approximately seven times as fast as *D. simulans.* These unequal rates are surprising and have not been previously described.

Evidence for a Gene 3′ of *Adh*

An unexpected finding is the apparent presence of a previously unidentified gene 125 bases downstream from the *Adh* poly A addition site. The distribution of A+T is consistent with this interpretation (fig. 3), and typical consensus sequences (TATA box, ATG, and splice junctions) occur at positions predicted by the A+T distribution. Further interspecies differences are limited to single base substitutions (mostly in the third codon position) in the putative exons, while in the putative introns there are also examples of length variation. Finally, among more distantly related species, the degree of observed sequence conservation in this region is consistent with the presence of coding regions (S. Schaeffer and R. Blackman, personal communication).

There are a number of other examples of clustered genes in *Drosophila,* including those coding for amylase (Levy et al. 1985), heat-shock proteins, chorion, yolk proteins, and larval cuticle proteins (reviewed by Sirotkin and Davidson 1982). Additionally, Henikoff et al. (1986) have identified a pupal cuticle-protein gene nested within an intron of a *Drosophila* purine-pathway gene. If only ∼10% of the genome codes for protein (Sirotkin and Davidson 1982), the expected random gene distribution would be ∼1 gene/8 kb. Thus, it is perhaps surprising to find so many examples of tightly clustered genes. Genetic or cytogenetic mapping has not revealed tight gene clustering, but these studies lack resolution on the molecular level. From the examples listed above, it appears that clustering of genes in *Drosophila* may be the rule rather than the exception.

Acknowledgments

We are grateful to Dr. D. Goldberg for providing us with the *Drosophila melanogaster Adh* clone and to Dr. M. Meselson for genomic libraries. We thank Drs. C. Benyajati, M. Kreitman, J. Posakony, and D. Sullivan for allowing us to discuss some of their findings in advance of publication and Dr. L. Grossman for assistance with the computer analysis. We particularly wish to thank Dr. W. Fitch for his helpful comments and for his assistance in constructing the phylogeny. This research was supported by National Institutes of Health (NIH) training grants GM-07315 and HD-07274 to V.H.C. and NIH grant GM-28851 to G.P.M.

LITERATURE CITED

ASHBURNER, M., M. BODMER, and C. LEMEUNIER. 1984. On the evolutionary relationships of *D. melanogaster.* Dev. Genet. **4**:295–312.

BATTERHAM, P., J. LOVETT, W. STARMER, and D. SULLIVAN. 1983. Differential regulation of duplicate alcohol dehydrogenase genes in *D. mojavensis.* Dev. Biol. **96**:346–354.

BENYAJATI, C., and J. DRAY. 1984. Cloned *Drosophila* alcohol dehydrogenase genes are correctly expressed after transfection into *Drosophila* cells in culture. Proc. Natl. Acad. Sci. USA **81**: 1701–1705.

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

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

BENYAJATI, C., N. WANG, A. REDDY, E. WEINBERG, and W. SOFER. 1980. Alcohol dehydrogenase in *Drosophila:* isolation and characterization of messenger RNA and cDNA clone. Nucleic Acids Res. **8**:5649–5667.

BIGGIN, M., T. GIBSON, and G. HONG. 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.

BODMER, M., and M. ASHBURNER. 1984. Conservation and change in the DNA sequences coding for *Adh* in sibling species of *Drosophila.* Nature **309**:425–430.

BREATHNACH, R., C. BENOIST, K. O'HARE, F. GANNON, and P. CHAMBON. 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. USA **75**:4853–4857.

CANN, R., and A. WILSON. 1983. Length mutations in human mitochondrial DNA. Genetics **104**:699–711.

COHN, V., M. THOMPSON, and G. MOORE. 1984. Nucleotide sequence comparison of the *Adh* gene in three drosophilids. J. Mol. Evol. **20**:31–37.

COYNE, J. 1984. Genetic basis of male sterility in hybrids between two closely related species of *Drosophila.* Proc. Natl. Acad. Sci. USA **81**:4444–4447.

DEININGER, P. 1983. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. **129**:216–223.

DICKINSON, W., R. ROWAN, and M. BRENNAN. 1984. Regulatory gene evolution: adaptive differences in expression of alcohol dehydrogenase in *D. melanogaster* and *D. simulans.* Heredity **52**:215–225.

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

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**: 279–284.

GOJOBORI, T., W. LI, and D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

GOLDBERG, D. 1980. Isolation and partial characterization of the *Drosophila* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA **77**:5794–5798.

GOLDBERG, D., J. POSAKONY, and T. MANIATIS. 1983. Correct developmental expression of a cloned alcohol dehydrogenase gene transduced into the *Drosophila* germ line. Cell **34**:59–73.

GOLDBERG, M. 1979. Sequence analysis of *Drosophila* histone genes. Ph.D. thesis, Stanford University, Palo Alto, Calif.

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

HENIKOFF, S., M. KEENE, K. FECHTEL, and J. FRISTOM. 1986. Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell **44:**33–42.

HUNT, J., T. HALL, and R. BRITTEN. 1981. Evolutionary distances in Hawaiian *Drosophila* measured by DNA reassociation. J. Mol. Evol. **17:**361–367.

KREITMAN, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *D. melanogaster.* Nature **304:**412–417.

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

LAIRD, C., and B. MCCARTHY. 1968. Magnitude of interspecific nucleotide sequence variability in *Drosophila.* Genetics **60:**303–322.

LEVY, J., R. GEMMILL, and W. DOANE. 1985. Molecular cloning of alpha-amylase genes from *D. melanogaster.* II. Clone organization and verification. Genetics **110:**313–324.

MOORE, G. 1984. The C-value paradox. BioScience **34:**425–429.

OAKESHOTT, J., G. CHAMBERS, P. EAST, J. GIBSON, and J. BARKER. 1982. Evidence for a genetic duplication involving *Adh* in *D. buzzatii* and related species. Aust. J. Biol. Sci. **35:** 73–84.

PERLER, F., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KOLODNER, and J. DODGSON. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20:**555–566.

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

SIROTKIN, K., and N. DAVIDSON. 1982. Developmentally regulated transcription from *D. melanogaster* chromosomal site 67B. Dev. Biol. **89:**196–210.

URSPRUNG, H., K. SMITH, W. SOFER, and D. SULLIVAN. 1968. Assay systems for the study of gene function. Science **160:**1075–1081.

URSPRUNG, H., W. SOFER, and N. BURROUGHS. 1970. Ontogeny and tissue distribution of alcohol dehydrogenase in *D. melanogaster.* Wilhelm Roux Arch. **164:**201–208.

ZWIEBEL, L., V. COHN, D. WRIGHT, and G. MOORE. 1982. Evolution of single-copy DNA and the *Adh* gene in seven drosophilids. J. Mol. Evol. **19:**62–71.

WALTER M. FITCH, reviewing editor

Received June 9, 1986; revision received November 4, 1987