Molecular Analysis of Duplicated Esterase Genes in *Drosophila melanogaster*[1]

*C. Collet,**,[2] *K. M. Nielsen,** *R. J. Russell,*† *M. Karl,** *J. G. Oakeshott,*† *and R. C. Richmond**

*Department of Biology and Institute for Molecular and Cellular Biology, Indiana University; and †CSIRO Division of Entomology, Canberra

Genomic clones containing sequences homologous to an esterase 6 (*Est-6*) cDNA clone were isolated from a library of *Drosophila melanogaster* DNA. Comparison of the genomic and cDNA sequences revealed that the *Est-6* gene comprises two exons, one of 1,387 bp and one of 248 bp, separated by a short intron of 51 bp. Further sequencing revealed the presence of a tandem duplication of the *Est-6* gene (denoted *Est-P*) which also has an exon of 1,387 bp and an exon of 248 bp, separated by a short intron of 56 bp. The two genes show similarities of 64% and 60% at the DNA and protein levels, respectively. The coding regions of the genes are 197 bases apart, and presumptive 5′ regulatory sequences of *Est-P* overlap at least the 3′ noncoding region of *Est-6*. Transcripts homologous to *Est-P* were detected in late larvae and adults of each sex, whereas *Est-6* transcripts are present in all life stages but are predominant in adult males. This suggests different physiological functions for the products of the two genes. Southern and Northern blot hybridization analyses of the 20-kb region surrounding the *Est-6*/*Est-P* duplication failed to detect any other duplicated esterase genes, although this region is actively transcribed.

Introduction

Comparison of the full cDNA sequences of several esterase enzymes of diverse origin and function suggests the existence of a multigene family of serine esterases that is distinct from the functionally related serine protease multigene family (Myers et al. 1988). The proposed serine esterase family presently contains several cholinesterases (Hall and Spierer 1986; Schumaker et al. 1986; McTiernan et al. 1987) and at least one carboxylesterase, EST 6, from *Drosophila melanogaster* (Oakeshott et al. 1987). These enzymes share ≥25% similarity in their amino acid sequence, and all contain a highly conserved octapeptide surrounding an invariant serine. Biochemical analysis has shown this serine residue to be directly involved in the hydrolysis of substrates (Dayhoff et al. 1972). The only eukaryotic esterase so far sequenced that does not lie within the serine esterase family is a human carboxylesterase, EST D. This enzyme has little functional or structural similarity either to the other esterases or to the serine proteases (Lee and Lee 1986).

Although less direct, genetic evidence for clusters of tightly linked carboxylesterases in mammals (Hedrich and von Deimling 1987) and *Drosophila* (Zouros et al. 1982) is also consistent with the existence of a serine esterase multigene family. Perhaps the

1. Key words: *Drosophila,* esterases, duplicate genes, divergent regulation.

2. Present address: CSIRO Division of Wildlife and Ecology, Canberra, A.C.T., 2602, Australia.

Address for correspondences and reprints: Dr. R. C. Richmond, Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, Indiana 47405.

Mol. Biol. Evol. 7(1):9–28. 1990.

0737-4038/90/0701-0002$02.00

best-characterized cluster involves two carboxylesterase loci in *D. mojavensis* and in its close relative *D. buzzatii*. These two loci, denoted *Est-4* and *Est-5* in *D. mojavensis*, have not been separated by genetic recombination (Zouros et al. 1982). Limited sequence analysis of the N-terminal amino acid residues of the EST 4 and EST 5 proteins has confirmed previous biochemical analyses showing that the two proteins are closely related structurally, both to each other and to EST 6 of *D. melanogaster* (Pen et al. 1986, and submitted). However, the patterns of expression of the two enzymes—and presumably their physiological functions—are qualitatively different from each other and from EST 6. EST 4 is largely confined to the cuticle of late larvae, and EST 5 is present in the hemolymph and fat body through much of the life cycle (Zouros et al. 1982), while EST 6 is also found in the hemolymph but is predominant in the anterior ejaculatory duct of the adult male (Sheehan et al. 1979). Thus, the carboxylesterases in *Drosophila* are a model system for the study of the evolution of functional differences between members of a multigene family through the processes of gene duplication and the subsequent divergence of regulatory sequence information.

Although there is no genetic evidence for a duplication of a carboxylesterase gene in *D. melanogaster*, the present paper presents molecular evidence for a carboxylesterase duplication in this species which appears to be homologous to that in *D. mojavensis*. Sequence data are presented for a region of the genome of *D. melanogaster* that encompasses *Est-6* and an adjacent open reading frame. This open reading frame, denoted *Est-P*, has essentially the same exon/intron structure as and 64% DNA sequence similarity to *Est-6*. However, comparisons between the 5′ flanking regions of the two genes reveal substantial differences, suggesting that the two genes may be regulated differently. Consistent with this proposal, developmental Northern blot analysis indicates that *Est-P* is mainly transcribed in late larvae. We therefore propose that *Est-6* and *Est-P* in *D. melanogaster* are homologous to the *Est-4*/*Est-5* duplication in *D. mojavensis* and, from genetic and biochemical evidence, to the *Est-1*/*Est-J* duplication in *D. buzzatii* (East 1984; Knibb et al. 1987). We further propose that *Est-P*, *Est-4*, and *Est-J* are homologous and relatively conserved in their regulation, while *Est-6*, *Est-5*, and *Est-1* are also homologous but divergent with respect to their regulation.

Material and Methods

Nucleic Acid Preparations

Genomic DNA from flies homozygous for *Est-6*^{S} (Dm145; Scott 1986), plasmid DNA, and bacteriophage lambda DNA were prepared and analyzed by standard procedures (Maniatis et al. 1982). Total cellular and poly-A$^+$ RNA from several different life stages of the Canton-S strain, also homozygous for Est 6^S, were prepared and analyzed according to a method described by Oakeshott et al. (1987).

Hybridizations

A library of Dm145 genomic DNA was constructed by ligating a partial *Sau*3A digest into λ EMBL4. The library was screened with the *Est-6* cDNA clone (Oakeshott et al. 1987) by using standard conditions (Maniatis et al. 1982). Double-stranded DNA probes were prepared by nick-translation (Rigby et al. 1977).

Southern blot hybridizations were performed using Zeta-probe membranes (Bio-Rad) and the alkali transfer procedure and hybridization conditions described by Reed and Mann (1985).

For Northern blot hybridizations, RNA was transferred to Zeta-probe membranes, prehybridized for 4 h, and hybridized overnight at 50°C in 1 × SSPE [0.18 M NaCl,

10 mM sodium phosphate, pH 7.0, 1 mM ethylenediaminetetraacetic acid (EDTA)], 1% sodium dodecyl sulfate (SDS), 0.5% skim milk powder, and 60% (v/v) formamide containing 0.5 mg carrier DNA/ml. Single-stranded RNA probes were transcribed, using either T7 or T3 RNA polymerase, from inserts cloned into the pBS M13$^+$ vector (Stratagene Cloning Systems). After hybridization, membranes were washed for 15 min at room temperature with 2 × SSC/0.1% SDS and for 1 h at 68°C with 0.2 × SSC (20 × SSC = 3 M NaCl, 0.3 M Na_3 citrate $2H_2O$, pH 7.0)/1% SDS and then were treated for 15 min at room temperature with 1 μg RNase A/ml in 2 × SSC and were washed with 0.2 × SSC/1% SDS for 30 min at 50°C.

Primer Extension and RNase Protection

RNA was isolated from individuals of the Dm145 strain. Material was ground in 0.2 M NaCl, 20 mM Tris-HCl, pH 7.5, 20 mM EDTA, 2% SDS. Proteinase-K (Sigma) was added to 250 μg/ml and incubated at 50°C for 45 min. Following two extractions with SEVAG (phenol:chloroform:isoamyl alcohol, 25:24:1), nucleic acids were precipitated with ethanol. The pellet was then digested with RQ1 DNase (Promega), extracted with SEVAG, and precipitated with ethanol. Poly-A$^+$ RNA was isolated according to a method described by Oakeshott et al. (1987).

For primer extension, 10^5 cpm of 5′-end-labeled oligonucleotide (5′-GGTCAT-CTGTATCACTCGC-3′) was annealed to 5 μg female or 1.5 μg male poly-A$^+$ RNA, to the total RNA isolated from ejaculatory ducts dissected from three 3–5-day-old males, or to 50 μg yeast tRNA (Sigma) as a control. For RNase protection, single-stranded RNA probes were transcribed from both strands of the 1.3-kb *Bam*HI fragment (bases 1193–2523). Samples (10^5 cpm) of probe were annealed at 50°C or 55°C to 5 or 20 μg of total RNA from 3–5-day-old males. Annealing, extension with reverse transcriptase, RNase protection, and electrophoretic conditions were done according to methods described by Ausubel et al. (1987, pp. 4.8.1–4.8.3).

DNA Sequencing

Both strands of the genomic clones were sequenced using M13mp18 and M13mp19 vectors (Norrander et al. 1983) and the dideoxy chain-termination method of Sanger et al. (1977).

Results

Est-6 Gene Structure

The nucleotide sequence was determined for 4.1 kb of DNA from a genomic clone that encompasses the *Est-6* gene (fig. 1). The EST 6 coding region is located from base 245 to base 1930, and comparison with the cDNA sequence (Oakeshott et al. 1987) reveals that it contains a single intron of 51 bp, from base 1632 to base 1682. This contrasts with *Drosophila* acetylcholinesterase (*Ace*), which may have at least two introns of considerable length (Hall and Spierer 1986).

Primer extension analysis was performed using a 20-mer oligonucleotide complementary to bases 305–324 in the *Est-6* sequence (fig. 2A). The results suggest that *Est-6* has multiple transcription initiation sites between base 204 and base 210. The relative intensities of the bands indicate that the guanine at base 207, adenine at base 210, and the thymidine at base 209 are the primary start sites, with the other bases used less frequently. The CAP-site region of the *Est-6* gene includes several nucleotides identified by Bucher and Trifonov (1986) as characteristic of eukaryote promoters. Furthermore, the short leader sequence of 35–41 bp of *Est-6* is typical of eukaryotic

```
                *         *         *         *         *         *         *         *         *         *
1   GTATACGGCTATCGTTTTAATTCGCACACGCCTATCAACTGGATGATGTTCACACTAGAGTTACTCCCATTTGAAAGCCGGGCATTGGAAAACTAATGTC

                *         *         *         *         *         *         *         *         *         *
101 ATGCGTGCCCAGATCTCAATTGAGACTGGTTGACTGGATGTTCAGGTGGCCGGGTGGCGATAAGCCGATCGATGGAATAAAAGGGGCCGCAATTGCCGCA

                *         *         *         *         *         *         *         *         *         *
201 TCTCAAGATAGTTGCGGTCTGAATTCGCCGGAGTGAGGAGCAACATGAACTACGTGGGACTGGGACTTATCATTGTGCTGAGCTGCCTTTGGCTCGGTTC
                                                    MetAsnTyrValGlyLeuGlyLeuIleIleValLeuSerCysLeuTrpLeuGlySe

                *         *         *         *         *         *         *         *         *         *
301 GAACGCGAGTGATACAGATGACCCTCTGTTGGTGCAGCTGCCCCAGGGCAAGCTACGTGGTCGCGATAATGGAAGCTACTACAGCTACGAATCGATTCCC
    rAsnAlaSerAspThrAspAspProLeuLeuValGlnLeuProGlnGlyLysLeuArgGlyArgAspAsnGlySerTyrTyrSerTyrGluSerIlePro

                *         *         *         *         *         *         *         *         *         *
401 TACGCCGAACCGCCCACTGGCGATCTACGATTCGAGGCTCCAGAGCCGTACAAACAAAAGTGGTCGGATATATTCGATGCCACCAAAACCCCGGTGGCGT
    TyrAlaGluProProThrGlyAspLeuArgPheGluAlaProGluProTyrLysGlnLysTrpSerAspIlePheAspAlaThrLysThrProValAlaC

                *         *         *         *         *         *         *         *         *         *
501 GCCTGCAGTGGGATCAGTTCACGCCTGGGGCCAACAAATTGGTAGGAGAGGAGGATTGCCTAACCGTCAGCGTCTACAAGCCGAAGAATAGCAAGAGGAA
    ysLeuGlnTrpAspGlnPheThrProGlyAlaAsnLysLeuValGlyGluGluAspCysLeuThrValSerValTyrLysProLysAsnSerLysArgAs

                *         *         *         *         *         *         *         *         *         *
601 TAGCTTTCCGGTGGTGGCCCACATTCACGGAGGTGCCTTTATGTTCGGTGCAGCATGGCAAAATGGACACGAGAACGTGATGCGTGAGGGCAAATTCATT
    nSerPheProValValAlaHisIleHisGlyGlyAlaPheMetPheGlyAlaAlaTrpGlnAsnGlyHisGluAsnValMetArgGluGlyLysPheIle

                *         *         *         *         *         *         *         *         *         *
701 CTGGTGAAGATAAGCTATCGCCTGGGGCCATTGGGTTTCGTGAGCACCGGCGATAGGGATCTTCCCGGAAACTATGGACTGAAAGATCAACGGCTGGGTC
    LeuValLysIleSerTyrArgLeuGlyProLeuGlyPheValSerThrGlyAspArgAspLeuProGlyAsnTyrGlyLeuLysAspGlnArgLeuAlaL
```

FIG. 1 (*Continued*)

```
                *         *         *         *         *         *         *         *         *         *
1501 GAATAGCACGCAGGAGTCATTGGATCTTCATCGCAAATATGGAAAGAGTCCTGCCTACGCTTATGTCTATGACAATCCAGCCGAAAAAGGAATCGCACAG
     sAsnSerThrGlnGluSerLeuAspLeuHisArgLysTyrGlyLysSerProAlaTyrAlaTyrValTyrAspAsnProAlaGluLysGlyIleAlaGln

                *         *         *         *         *         *         *         *         *         *
1601 GTCCTGGCCAATCGAACCGATTATGATTTTGGTAAGGAAATCGTACTTTTAAATGGACTTAGTTAAATCATTACTTTTATAGGAACTGTACACGGTGACG
     ValLeuAlaAsnArgThrAspTyrAspPheG                                                   lyThrValHisGlyAspA

                *         *         *         *         *         *         *         *         *         *
1701 ACTACTTTTTGATATTCGAAAATTTCGTACGAGATGTGGAAATGCGTCCGGATGAGCAGATAATTTCGAGAAATTTTATCAATATGCTGGCAGATTTGGC
     spTyrPheLeuIlePheGluAsnPheValArgAspValGluMetArgProAspGluGlnIleIleSerArgAsnPheIleAsnMetLeuAlaAspPheAl

                *         *         *         *         *         *         *         *         *         *
1801 TTCGAGTGATAATGGCTCTCTAAAATATGGTGAATGCGATTTCAAAGATAGTGTAGGTAGTGAGAAATTCCAATTATTAGCTATTTATATTGATGGCGGC
     aSerSerAspAsnGlySerLeuLysTyrGlyGluCysAspPheLysAspSerValGlySerGluLysPheGlnLeuLeuAlaIleTyrIleAspGlyGly

                *         *         *         *         *         *         *         *         *         *
1901 CAGAATAGGCAGCATGTGGAATTTCCGTAAGTTACATGAATAAAATCAAAAATTTTTCGTTCTGTGTAATTTTTAATTATTTAATTATTCTCAACTGGCT
     GlnAsnArgGlnHisValGluPhePro

                *         *         *         *         *         *         *         *         *         *
2001 TTAAATATCATTTGTACAAAACGTGTTTGTCGCTTTATATTTTGGTTTTTGTTGTTTCTTTATAAAGAATAATAAACGTTTGTTTACCCGCAGTTTGATT

                *         *         *         *         *         *         *         *         *         *
2101 AGTACTTTGTGGAGTTCCAGATCATGAGTATATTCAAACGGCTGTTGTGCCTGACTTTGCTGTGGATAGCAGCTTTAGAATCTGAAGCTGATCCCTTGAT
                          MetSerIlePheLysArgLeuLeuCysLeuThrLeuLeuTrpIleAlaAlaLeuGluSerGluAlaAspProLeuIl

                *         *         *         *         *         *         *         *         *         *
2201 TGTTGAGATAACAAATGGAAAAATCCGTGGCAAAGATAATGGGTTGTACTACAGCTACGAATCGATTCCCTATGCCGAGCATCCAACTGGTGCCCTCCGT
     eValGluIleThrAsnGlyLysIleArgGlyLysAspAsnGlyLeuTyrTyrSerTyrGluSerIleProTyrAlaGluHisProThrGlyAlaLeuArg
```

```
               *         *         *         *         *         *         *         *         *         *
2301 TTTGAAGCACCTCAGCCGTATAGTCATCATTGGACTGATGTTTTCAATGCCACGCAGTCTCCAGTTGAGTGCATGCAGTGGAATCAGTTTATAAACGAAA
     PheGluAlaProGlnProTyrSerHisHisTrpThrAspValPheAsnAlaThrGlnSerProValGluCysMetGlnTrpAsnGlnPheIleAsnGluA

               *         *         *         *         *         *         *         *         *         *
2401 ACAATAAGCTGATGGGTGATGAGGATTGCTTAACGGTAAGCATCTATAAGCCAAAGAAACCCAATCGGAGCAGCTTTCCTGTCGTAGTACTCCTGCATGG
     snAsnLysLeuMetGlyAspGluAspCysLeuThrValSerIleTyrLysProLysLysProAsnArgSerSerPheProValValValLeuLeuHisGl

               *         *         *         *         *         *         *         *         *         *
2501 AGGTGCTTTCATGTTCGGTAGTGGATCCATATATGGACACGACTCCATTATGCGTGAGGGAACTTTGCTTGTGGTAAAAATAAGCTATCGTCTTGGACCA
     yGlyAlaPheMetPheGlySerGlySerIleTyrGlyHisAspSerIleMetArgGluGlyThrLeuLeuValValLysIleSerTyrArgLeuGlyPro

               *         *         *         *         *         *         *         *         *         *
2601 TTGGGTTTTGCAAGTACCGGCGATAGACACTTGCCGGGAAACTATGGTCTAAAGGATCAACGTCTGGCCCTACAATGGATCAAGAAGAACATTGCTCACT
     LeuGlyPheAlaSerThrGlyAspArgHisLeuProGlyAsnTyrGlyLeuLysAspGlnArgLeuAlaLeuGlnTrpIleLysLysAsnIleAlaHisP

               *         *         *         *         *         *         *         *         *         *
2701 TTGGTGGAATGCCAGATAATATTGTGCTCATTGGTCACTCTGCAGGCGGTGCTTCGGCTCATTTGCAGCTGTTGCACGAGGATTTCAAACATTTGGCCAA
     heGlyGlyMetProAspAsnIleValLeuIleGlyHisSerAlaGlyGlyAlaSerAlaHisLeuGlnLeuLeuHisGluAspPheLysHisLeuAlaLy

               *         *         *         *         *         *         *         *         *         *
2801 AGGAGCGATTTCGGTGAGCGGCAATGCATTGGATCCTTGGGTCATACAGCAGGGTGGACGACGACGTGCATTTGAACTGGGTCGTATTGTCGGTTGTGGA
     sGlyAlaIleSerValSerGlyAsnAlaLeuAspProTrpValIleGlnGlnGlyGlyArgArgArgAlaPheGluLeuGlyArgIleValGlyCysGly
```

FIG. 1 (*Continued*)

```
                *         *         *         *         *         *         *         *         *         *
2901 CACACAAATGTCTCCGCAGAACTCAAGGACTGCTTGAAGTCTAAGCCGGCTAGCGATATAGTCTCTGCTGTCCGAAGCTTCCTTGTGTTTTCCTATGAAC
     HisThrAsnValSerAlaGluLeuLysAspCysLeuLysSerLysProAlaSerAspIleValSerAlaValArgSerPheLeuValPheSerTyrValP

                *         *         *         *         *         *         *         *         *         *
3001 CCTTCAGTGCTTTTGGACCTGTTGTGGAGCCGTCAGATGCACCAGACGCCTTTCTAACCGAGGACCCAAGAGCAGTGATTAAGAGCGGGAAGTTTGCCCA
     roPheSerAlaPheGlyProValValGluProSerAspAlaProAspAlaPheLeuThrGluAspProArgAlaValIleLysSerGlyLysPheAlaGl

                *         *         *         *         *         *         *         *         *         *
3101 AGTCCCTTGGGCTGTGACGTACACCACTGAGGACGGGGGATACAACGCTGCTCAGCTGTTGGAAAGAAACAAATTAACTGGCGAGAGTTGGATTGACCTA
     nValProTrpAlaValThrTyrThrThrGluAspGlyGlyTyrAsnAlaAlaGlnLeuLeuGluArgAsnLysLeuThrGlyGluSerTrpIleAspLeu

                *         *         *         *         *         *         *         *         *         *
3201 CTCAATGATCGATGGTTTGATATATGGATGTACTTGCTCTTCTATCGGGACGCCAAGAAAACCATCAAAGATATGGATGATCTTTCATTTGATCTCAGGC
     LeuAsnAspArgTrpPheAspTrpAlaProTyrLeuLeuPheTyrArgAspAlaLysLysThrIleLysAspMetAspAspLeuSerPheAspLeuArgG

                *         *         *         *         *         *         *         *         *         *
3301 AGCAGTATCTAGCAGATCGGCGATTCAGTGTGGAAAGTTATTGGAACGTGCAGCGAATGTTTACTGATGTTCTTTTCAAGAATAGCGTGCCAAGTGCAAT
     lnGlnTyrLeuAlaAspArgArgPheSerValGluSerTyrTrpAsnValGlnArgMetPheThrAspValLeuPheLysAsnSerValProSerAlaIl

                *         *         *         *         *         *         *         *         *         *
3401 AGATCTTCACCGAAAGTATGGCAAAAGTCCGGTTTATTCTTTTGTCTACGATAATCCTACCGATTCCGGAGTGGGTCAATTGCTTTCCAATCGAACACGAT
     eAspLeuHisArgLysTyrGlyLysSerProValTyrSerPheValTyrAspAsnProThrAspSerGlyValGlyGlnLeuLeuSerAsnArgThrAsp

                *         *         *         *         *         *         *         *         *         *
3501 GTACATTTTGGTGGGTACATACTTGCGTTTCAAATACTTAAAGCTCTAGCTTTCTTATTATTAAAGGTACTGTCCACGGAGATGACTTTTTCTTGATTTT
     ValHisPheG                                                                    lyThrValHisGlyAspAspPhePheLeuIlePh
```

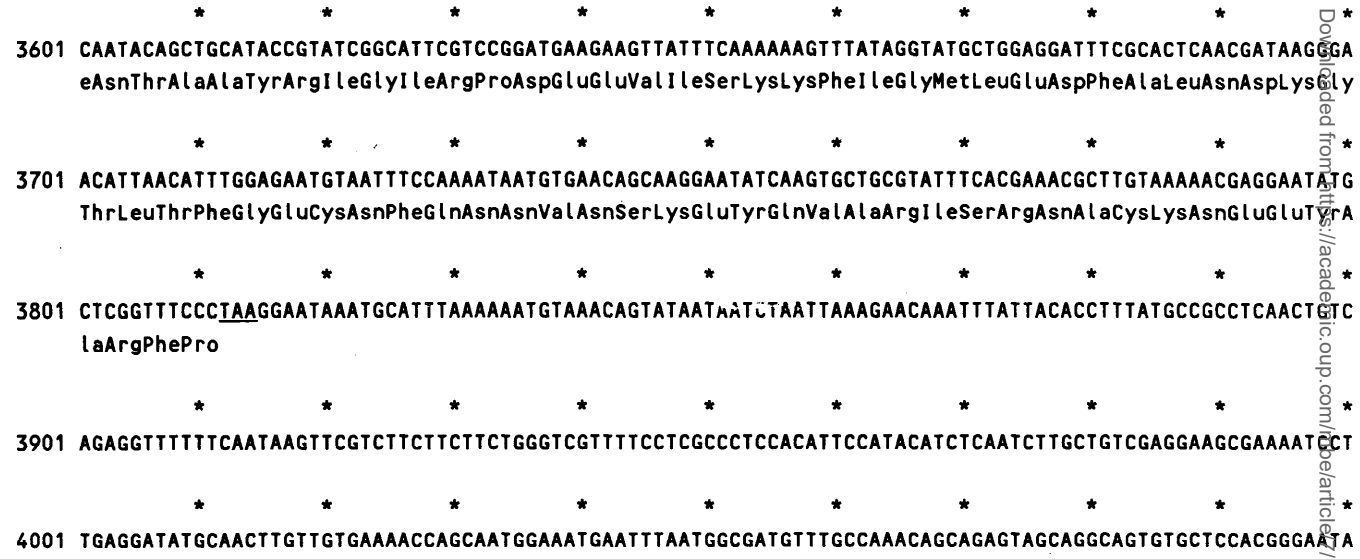

FIG. 1.—Sequence of 4,100 bp spanning the est 6 genomic region. The *Est-6* coding region is located between bases 245 and 1927, and the *Est-P* coding region is located between bases 2124 and 3811. The *Est-6* putative TATA box (AATAAAA) is around base 176, a CCAAT box is around base 120, a GC box is around base 152, and two polyadenylation signals are around bases 1939 and 2071. The genomic sequence presented here corrects the cDNA and protein sequence published by Oakeshott et al. (1987) and Myers et al. (1988). The inclusion of a base omitted from the cDNA sequence at position 1895 alters the amino acid sequence of the last 15 residues and shortens this to 11 residues (also see fig. 4).

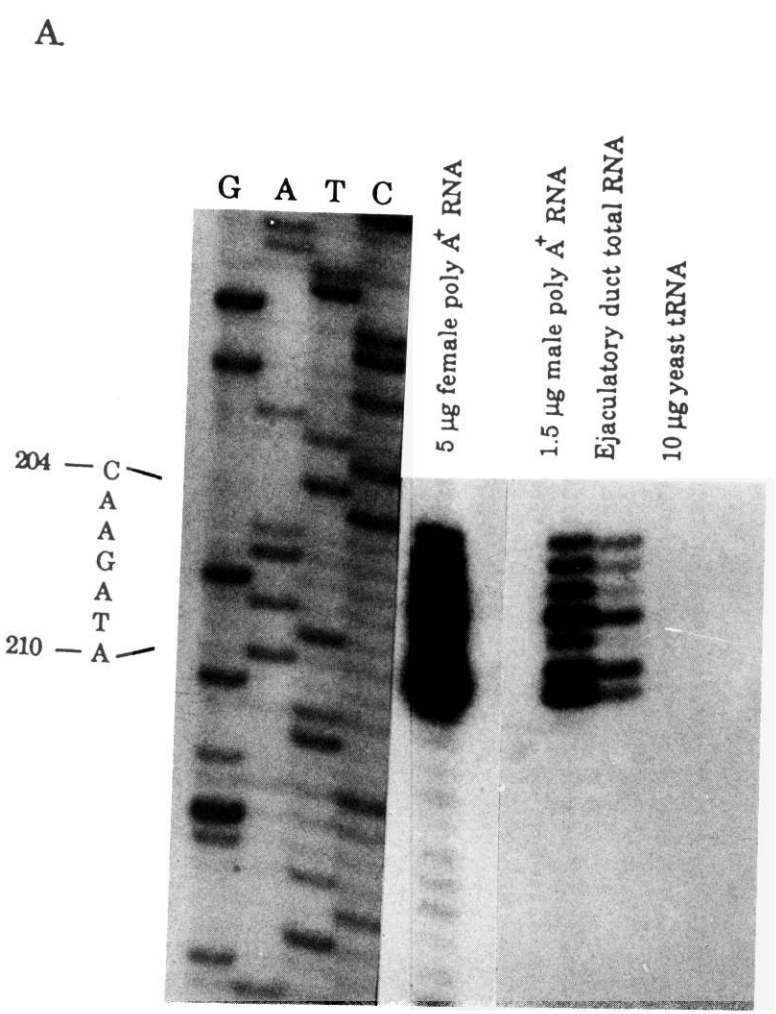

FIG. 2.—Determination of the transcription start site and polyadenylation sites of *Est-6* mRNA by primer extension and RNase protection. A, left four lanes, Sequencing ladder of an M13 subclone of the 5′ region of *Est-6* obtained by extending a labeled primer complementary to bases 305–324 (fig. 1). This primer was annealed to poly-A^+ RNA isolated from males and females or to total RNA isolated from dissected anterior ejaculatory ducts of males. The primer was also annealed to yeast tRNA as a control. A, right four lanes, Primer extended and resultant fragment sized on a standard sequencing gel. mRNA initiation sites are observed between bases 204 and 210. B, Map of the *Est-6* subclone used to produce cRNA for protecting the 3′ termini of the *Est-6* messages. The position of the intron, stop codon, and potential polyadenylation signal sites are shown. Fragment sizes indicated by double arrows below the map indicate bands expected following RNase digestion of single-stranded RNA if both of the potential polyadenylation signals are functional. C, RNase protection of *Est-6* cRNA transcribed from either the T3 (not complementary to *Est-6* mRNA) or T7 (complementary to *Est-6* mRNA) polymerase promoters which flank the 1.3-kb *Bam*HI subclone of *Est-6* shown in panel B. The two outside lanes show size markers. Total male RNA was annealed to either the T3- or T7-derived probe and subjected to RNase treatment. Bands corresponding to the fragments predicted in panel B are protected by the T7 probe.

leader sequences (Kozak 1983). However, it contrasts with the *Ace* gene of *Drosophila* and *Torpedo,* which have unusually long 5′ untranslated sequences containing multiple initiation and termination codons (Schumaker et al. 1986; Sikarov et al. 1987).

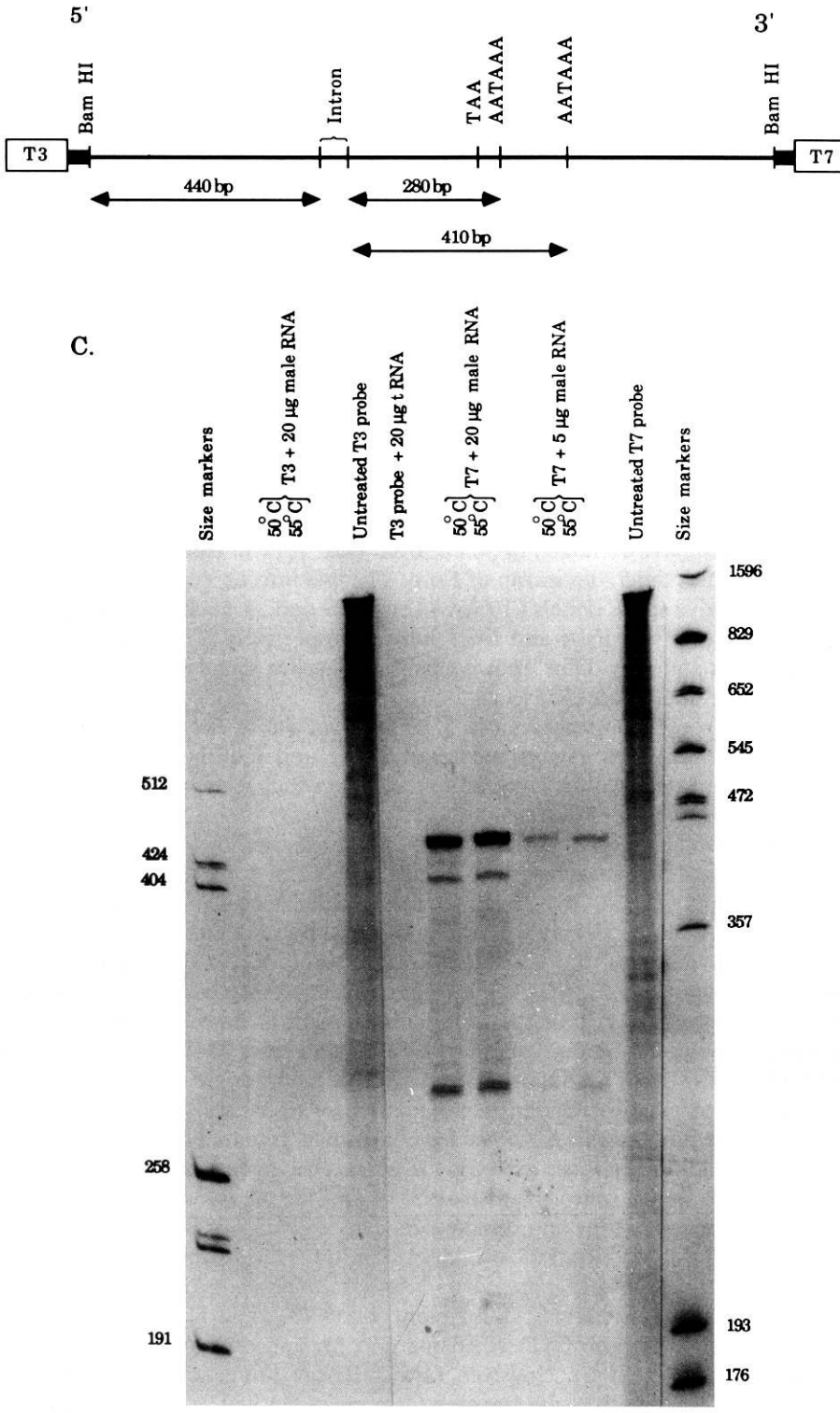

FIG. 2. (*Continued*)

The sequence around position 176 (AATAAAA) is most likely to be the TATA box of *Est-6*. It deviates from consensus in the first base but otherwise corresponds in position and composition to the TATA box region identified in many eukaryotic promoters (Bucher and Trifonov 1986). A potential CCAAT box is present around base 120, and a GC box around base 152 (Maniatis et al. 1987).

Two consensus polyadenylation signals (AATAAA; Wickens and Stephenson 1984) are present around bases 1939 and 2071. Use of these two sites would explain the results of Northern analysis (Oakeshott et al. 1987) showing two *Est-6* transcripts, 1.68 and 1.83 kb in length. RNase protection of *Est-6* cRNA confirms that both polyadenylation signals are used (fig. 2B). Protected bands correspond to sequences between the central *Bam*HI site (base 1192) and the 5′ end of the intron (base 1632) and from the 3′ end of the intron to around bases 1960 and 2100.

Duplication of an Esterase Gene

An additional open reading frame (*Est-P*) begins 197 bases 3′ of the *Est-6* termination codon and extends from base 2124 to base 3811 (fig. 1). The *Est-6* and *Est-P* genes appear to be the result of a tandem duplication, since they include regions with 66% nucleotide similarity in the first exon and 57% similarity in the second. *Est-P* also appears to have an intron of 56 bp, from base 3511 to base 3566, in the same relative position as the 51-bp intron of *Est-6*. The two introns share ~50% sequence similarity. Putative splice signals (TTAAA) begin 29 and 24 nucleotides upstream of the acceptor sites of the *Est-6* and *Est-P* introns, respectively. These sites are in the correct position and only differ from the consensus splice signal sequence in the last position (Keller and Noon 1985).

There are several possible regions 5′ of the start site of *Est-P* that could act as regulatory elements. These regions are found both 5′ and 3′ of the stop codon of *Est-6*. Further experiments will be required to elucidate which if any of these regions is/are the primary promoter elements of *Est-P*.

Developmental Northern blot analysis using a single-stranded RNA probe derived from a *Cla*I/*Aha*III fragment (bases 2263–3829) specific for *Est-P* revealed the presence of two transcripts (fig. 3). A 1.92-kb transcript was detected in late larvae, with a second, less abundant 1.25-kb transcript in late larvae and adults.

We hypothesize that the 1.92-kb mRNA is transcribed from the *Est-P* DNA sequence homologous to *Est-6* and utilizes either the consensus polyadenylation signal (AATAAA) immediately 3′ of the termination codon at base 3817 or alternative nonconsensus polyadenylation signals further 3′ around bases 3848 (AATAAT), 3856 (AATTAA), 3857 (ATTAAA), or 4038 (AATGAA). The origin of the 1.25-kb transcript is more problematical. On the basis of a search for sequences with strong similarity to the splice-site junctions in *Est-6*, we have identified potential donor and acceptor splice sites in addition to those used for the generation of the 1.92-kb message. Use of an alternative donor splice site at base 2814 and the acceptor splice site at 3566 would yield a message of the appropriate size. This scheme would result in an intron of 751 bases being spliced from the primary transcript. This model for the production of the 1.25-kb transcript is supported by Northern analysis using a single-stranded RNA probe from the *Bam*HI/*Bgl*II fragment (bases 2832–3401) specific for the putative large intron. This probe detected the 1.92-kb transcript but not the 1.25-kb transcript (data not shown). However, further RNase protection experiments and analysis of cDNA clones from the two message populations will be required to confirm the origins of the two *Est-P* messages.

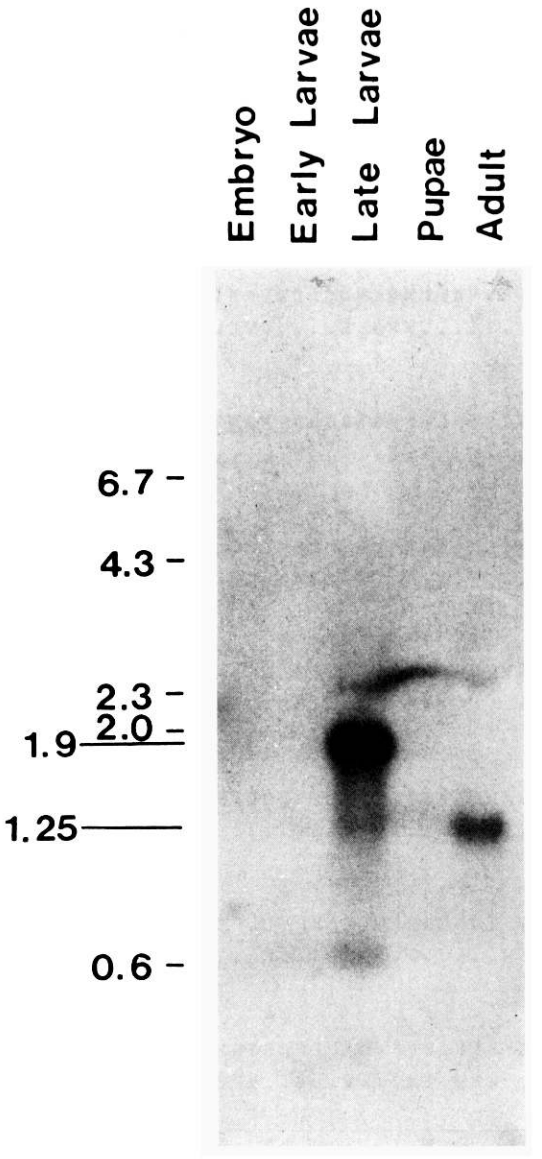

FIG. 3.—Northern blot hybridization analysis of the *Est-P* gene. A single-stranded RNA probe derived from a *Cla*I/*Aha*III fragment specific for *Est-P* was hybridized to a Northern blot containing 5 μg poly-A$^+$ RNA prepared from different life stages of Canton S. Molecular size markers are indicated.

Comparison of EST 6 and EST P Proteins

Conceptual translations of the *Est-6* and 1.92-kb *Est-P* transcripts yield proteins of 544 amino acids (fig. 4). The first 21 and 19 residues, respectively, of the inferred polypeptides probably define signal peptides, resulting in mature proteins of 523 and 525 amino acids for EST 6 and EST P, respectively. Sequence similarity between the two mature peptides is 66% in the region encoded by the first exon and 59% in the second exon region. Although there is low sequence similarity between the 21- and 19-residue N-terminal peptides, they both have properties consistent with signal pep-

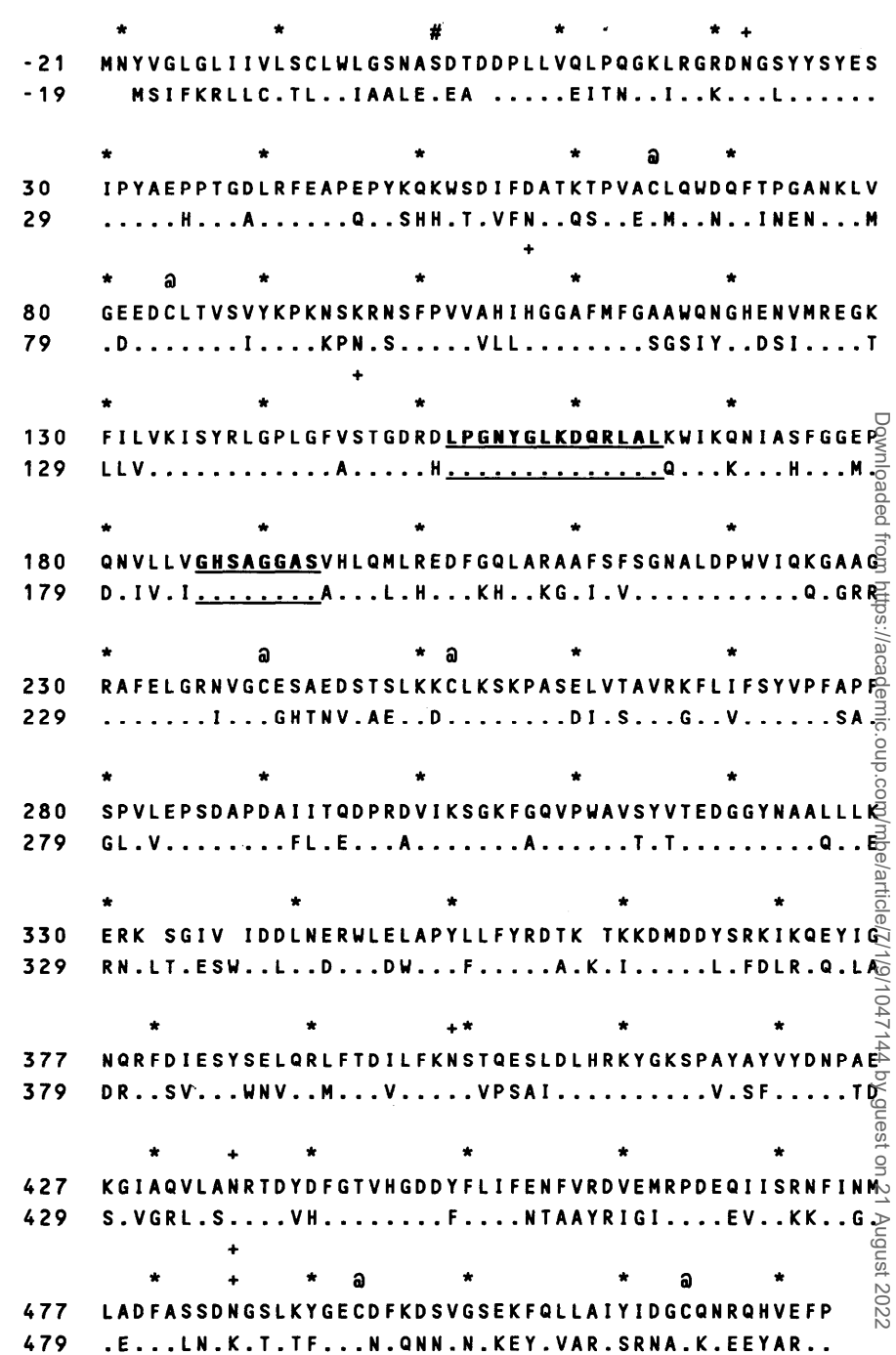

FIG. 4.—Comparison of the protein sequences of EST 6 and EST P. Numbers are relative to the first amino acid of the mature EST 6 protein, denoted by a pound sign (#). Regions of putative function are underlined. The first 21 and 19 amino acid residues of EST 6 and EST P, respectively, have properties consistent with signal peptides. The reactive serine is located at residue 188, and the Asp region implicated in proton transfer is around residue 160. Potential N-linked glycosylation sites are denoted by a plus sign (+), and cysteines are denoted by a circled "a" (@).

tides (Carne and Scheele 1982; Sjostrom et al. 1987). Tryptic peptide sequence analysis of the EST 6 protein also suggests that the first 21 residues of the inferred amino acid sequence are not retained in the mature protein (Oakeshott et al. 1987).

Each inferred amino acid sequence contains four potential N-linked glycosylation sites, at residues 21, 399, 435, and 485 for EST 6 and at residues 56, 95, 243, and 437 for EST P. Only one site, that at residue 435/437, is shared by both sequences. Most esterases studied to date have six cysteine residues (MacPhee-Quigley et al. 1986; Myers et al. 1988), and there are six conserved cysteines in both EST 6 and EST P (residues 65, 84, 240, 252, 493, and 514 in EST 6). Each cysteine residue in EST 6 must be involved in the formation of disulfide bridges, as there are no free thiol groups in the mature protein (Mane et al. 1983).

Structural and mutational analyses of several serine proteases have revealed that the catalytic mechanism involves three key residues which take part in a charge relay system to transfer a proton to the peptide bond of the substrate. The three residues—histidine, aspartate, and the reactive serine—are noncontiguous in the primary sequence, and each is surrounded by highly conserved regions of 8–18 residues (Price and Stevens 1982, p. 79; Craik et al. 1987; Carter and Wells 1988). The positions of analogous Asp- and Ser-containing regions in several esterases have been inferred by Sikarov et al. (1987) and Myers et al. (1988). The Asp-containing dodecapeptide and the Ser-containing octapeptide are around positions 160 and 188, respectively, in the EST 6/EST P alignment (fig. 4) and are absolutely conserved between the two proteins. Neither Sikarov et al. (1987) nor Myers et al. (1988) were able to identify a sequence in the esterases that was unambiguously analogous to the His-containing region involved in the catalytic mechanism of the serine proteases.

If the 1.25-kb *Est-P* transcript arises from differential splicing as indicated above, then it would encode an open reading frame of 240 codons that terminates with a TGA codon. The resultant polypeptide would contain some of the functionally constrained regions, including the reactive serine and aspartate residues implicated in proton transfer. However, it would only contain the two potential glycosylation sites and the two cysteine residues closest to the N-terminus.

Hybridization Analysis of the Region Surrounding the Esterase Duplication

To investigate the possibility that the *Est-6*/*Est-P* duplication is part of a larger cluster of esterase genes, similar to the two clusters of esterases in the rat (Hedrich and von Diemling 1987), 20 kb of genomic DNA surrounding the duplication were subjected to DNA and RNA blot hybridization analyses (fig. 5).

Low-stringency (50°C rather than 65°C) Southern blot hybridization analysis (data not shown) revealed that only fragments containing the *Est-6* or *Est-P* sequences hybridized to gel-purified *Est-6* or *Est-P* probes (see fig. 5D). These data suggest that the region from ~5 kb upstream to ~12 kb downstream of the *Est-6*/*Est-P* duplication contains no other esterase genes as similar to *Est-6* or *Est-P* as each is to the other.

However, this 20-kb region around the duplication is actively transcribed. Northern blot analysis revealed the presence of at least eight discrete size classes of developmentally specific poly-A^+ transcripts (fig. 5F). The locations of the sequences hybridizing to these transcripts suggest the existence of at least six different transcriptional units. Some of these transcriptional units may correspond to two lethal complementation groups, *l(3)69Ac* and *l(3)69Ad,* or to a visible mutation, *approximate* (*app*), which have been mapped close to *Est-6* at 69A1-5 (Hoogwerf et al. 1988). However,

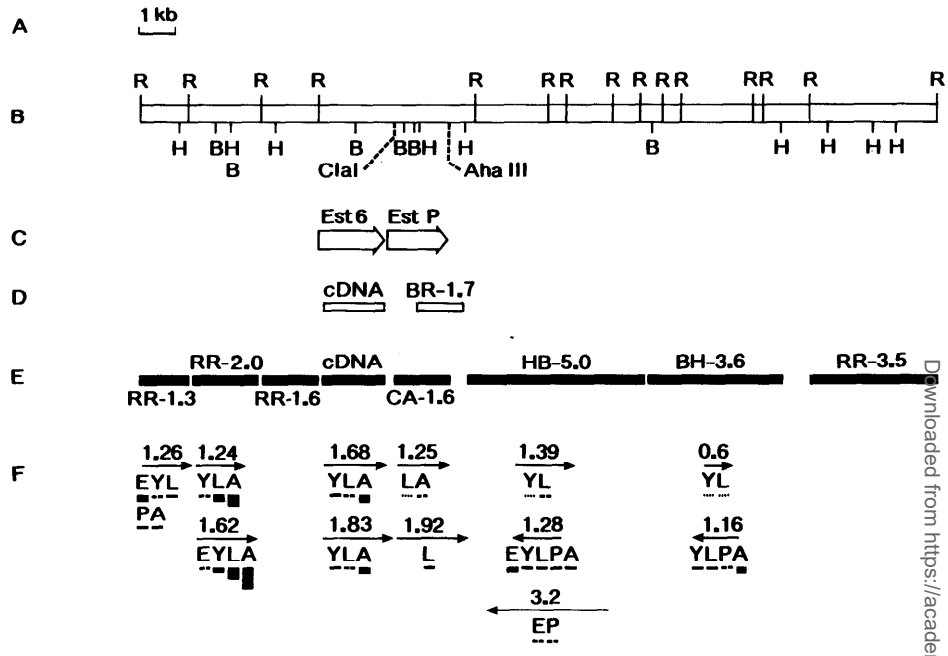

FIG. 5. Molecular organization of the genomic region around the *Est-6*/*Est-P* duplication. A, Scale in kilobases. B, Restriction map showing the positions of cleavage sites for *Eco*RI (R), *Hin*dIII (H), and *Bam*HI (B). The *Cla*I and *Aha*III sites which delineate an *Est-P*-specific fragment (see panel E below) are indicated by dashed lines. C, Positions of the *Est-6* and *Est-P* genes. Arrows refer to the direction of transcription. D, Fragments used to probe Southern blots containing cloned cDNAs which span the genomic region (see text). E, Fragments used for the production of single-stranded RNA probes for hybridization to Northern blots containing 5 μg poly-A^+ RNA prepared from five different life stages (see text). F, Summary of the Northern blot hybridization analysis. Each transcript and its direction of transcription are indicated by an arrow. The sizes of the transcript are given above the arrows, and life stages during which the transcripts were observed are given below the arrow (E = embryo; Y = early larvae; L = late larvae; P = pupae; A = adults). The relative intensities of the hybridization signals at each life stage are indicated.

both the orientation of the cloned region with respect to the chromosome and its precise relationships to these genes are unknown.

Discussion

DNA sequencing has revealed the presence of another putative esterase gene (*Est-P*) beginning 197 bases 3′ of the *Est-6* termination codon. Several lines of evidence suggest that *Est-P* is a functional gene: (1) Transcripts are produced from the *Est-P* gene, although at life stages different from those produced from the *Est-6* gene. (2) Splicing sites and mRNA processing signals appear to be intact; initiation and termination codons are present, and there are no premature termination codons. (3) Comparison of the *Est-6* and *Est-P* coding regions reveals a substantially smaller proportion of replacement-site (282/1,272 = 0.22) than silent-site differences (265/360 = 0.74), further suggesting functional constraint against mutations affecting the products of the two genes.

The predicted EST P polypeptide derived from the 1.92-kb transcript has 60% similarity to EST 6. Greater levels of protein similarity occur around some regions that have putative function. All six cysteines implicated in disulfide bonds in EST 6

are also found in EST P. The regions around the active-site serine (residues 186–193) and around the aspartate (residues 152–165) implicated in proton transfer during catalysis (Myers et al. 1988) are also conserved across the two proteins. The conservation of such regions suggests that the putative EST P polypeptide has catalytic activities similar to those of EST 6.

Although the *Est-P* and *Est-6* coding regions show 64% similarity, this similarity does not extend to the intron (46%) or to the putative promoter and 3′ untranslated regions, where the levels of similarity are insufficient for clear alignment. The lack of similarity over these regions may reflect either a lack of selective constraint or, particularly for the promoter regions, the different regulatory properties of the two genes.

The promoter sequences of *Est-P* will probably overlap the 3′ untranslated region and possibly even the coding sequence of *Est-6*. For example, the putative TATA box of *Est-P* (around 2061) is located upstream of the polyadenylation signal of *Est-6*. A similar situation may exist for alcohol dehydrogenase (*Adh*) and for a 3′ duplication in *Drosophila* (Schaeffer and Aquadro 1987), where the putative TATA box of the duplicated gene is 50 bp downstream of the polyadenylation signal of *Adh*. Elucidation of the extent of overlap between *Est-P* and *Est-6* and between the *Est-P* regulatory region and *Est-6* requires further analysis. However, one implication of the evidence to date is that the 5′ regulatory elements of *Est-P* may be under additional selective constraint if they correspond to transcribed and possibly translated regions of *Est-6*. In contrast, there is no such evidence for additional constraint on the *Est-6* promoter, since no other transcriptional activity was detected within 1.6 kb 5′ of the *Est-6* coding region.

We suggest that the *Est-6*/*Est-P* duplication is homologous to the *Est-4*/*Est-5* duplication of *D. mojavensis* (Pen et al., submitted) and to the *Est-1*/*Est-J* duplication of *D. buzzatii* (East 1984). In each case, the esterase genes involved in the duplication are on homologous chromosomes and in the same relative position to another linked esterase gene (chromosome 3 and *Est-C* in *D. melanogaster,* chromosome 2 and *Est-2* in *D. mojavensis* and *D. buzzatii;* Knibb et al. 1987). Furthermore, the N-terminal 50 amino acid residues of the mature EST 4 and EST 5 proteins display ~65% similarity to EST 6 and EST P (Pen et al., submitted). *Est-P, Est-4,* and *Est-J* all show similar developmental profiles with peak expression during late-larval stages (Zouros et al. 1982; East 1984). However, *Est-6, Est-5,* and *Est-1* differ in temporal and spatial aspects of their expression: EST 6 is found primarily in the adult male ejaculatory duct and to a lesser extent in the hemolymph (Sheehan et al. 1979); EST 5 is found in high concentration in the hemolymph and fat body (Zouros et al. 1982); and EST 1 is found in the hemolymph throughout the insect's life cycle (East 1984).

The conservation of developmental expression for esterases-P, -4, and -J—in comparison to the more variable expression of esterases-6, -5, and -1—would suggest a critical function for the former. Indeed, Zouros et al. (1982) have localized EST 4 to the late-larval cuticle, so this function may relate to changes in the cuticle structure preparatory to pupariation. However, the greater regulatory conservation of esterases-P, -4, and -J—relative to those of esterases-6, -5, and -1—may also reflect the additional constraint placed on the regulatory sequences of the former because of their possible location within the coding sequences of the latter.

The apparent homology between the *Est-6*/*Est-P* duplication and esterase genes found in *D. mojavensis* and *D. buzzatii* suggests that this duplication event is probably as old as the divergence of the *repleta* and *melanogaster* group. Throckmorton and others (Throckmorton 1975; Beverly and Wilson 1984; Blackman and Meselson 1986)

have estimated that the split between these two groups occurred 60–80 Myr ago (Mya). On the other hand, calculations of the age of the duplication event that are based on a comparison of nucleotide substitutions between *Est-P* and *Est-6* yield values of only 17–34 Mya, depending on the estimates of substitution rate used (Zweibel et al. 1982; Powell et al. 1986). However, the latter calculations are dubious because the estimates of substitution rates are derived from interspecific comparisons and do not account for phenomena such as gene conversion, which might occur between closely related genes in the same species.

Nevertheless, the present data are sufficient to indicate the promise of these duplicated carboxylesterases in *Drosophila* as models for the molecular analysis of long-held theories, not only about the evolution of new functions through gene duplication but also about the relative roles of regulatory and structural change in producing these new functions.

Acknowledgments

The authors wish to thank the following people for advice, assistance, enthusiasm, and critical review of the manuscript: S. Collet, A. Y. Game, M. Myers, M. J. Healy, J. Karotam, E. A. van Papenrecht, R. Joseph, J. Brady, P. H. Cooke, E. Snella, B. Nichols, and D. Sullivan. This work has been supported, in part, by NIH grant GM 39044 and NSF grant BSR-8718676.

LITERATURE CITED

AUSUBEL, F. R., R. BRENT, R. KINGSTON, D. MOORE, J. SEIDMAN, J. SMITH, and K. STRUHL. 1987. Current protocols in molecular biology. John Wiley & Sons, New York.

BEVERLY, S. M., and A. C. WILSON. 1984. Molecular evolution in *Drosophila* and the higher *Diptera*. II. A time scale for fly evolution. J. Mol. Evol. **21**:1–13.

BLACKMAN, R. K., and M. MESELSON. 1986. Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila* hsp82 gene. J. Mol. Biol. **188**:499–515.

BUCHER, P., and E. N. TRIFONOV. 1986. Compilation and analysis of eukaryotic POL II promoter sequences. Nucleic Acids Res. **14**:10009–10026.

CARNE, T., and G. SCHEELE. 1982. Amino acid sequences of transport peptides associated with canine exocrine pancreatic proteins. J. Biol. Chem. **257**:4133–4140.

CARTER, P., and J. A. WELLS. 1988. Dissecting the catalytic triad of a serine protease. Nature **332**:564–568.

CRAIK, C. S., S. ROCZNIAK, C. LARGMANN, and W. J. RUTTER. 1987. The catalytic role of active site aspartic acid in serine proteases. Science **237**:909–913.

DAYHOFF, M. O., W. C. BARKER, and J. K. HARDMAN. 1972. Evolutionary conservatism: enzyme structure and function. Pp. 89–99 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. **5**. National Biomedical Research Foundation, Washington, D.C.

EAST, P. D. 1984. Biochemical genetics of two highly polymorphic esterases in *Drosophila buzzatii*. Ph.D. thesis, University of New England, Armidale.

HALL, L. M. C., and P. SPIERER. 1986. The *Ace* locus of *Drosophila melanogaster:* structural gene for acetylcholinesterase with an unusual 5′ leader. EMBO J. **5**:2940–2954.

HEDRICH, H. J., and O. VON DIEMLING. 1987. Reevaluation of LG V of the rat and assignment of 12 carboxylesterases to two gene clusters. J. Hered. **78**:92–96.

HOOGWERF, A. M., M. AKAM, and D. ROBERTS. 1988. A genetic analysis of the *rose-gespleten* region (68C8-69B5) of *Drosophila melanogaster*. Genetics **118**:665–670.

KELLER, E. B., and W. A. NOON. 1985. Intron splicing: a conserved internal signal in introns of *Drosophila* pre-mRNAs. Nucleic Acids Res. **13**:4971–4981.

KNIBB, W. R., P. D. EAST, and J. S. F. BARKER. 1987. The polymorphic chromosome inversion and esterase complex in *Drosophila buzzatii*. I. Linkage disequilibria. Aust. J. Biol. Sci. **40:** 257–269.

KOZAK, M. 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles. Microbiol. Rev. **47:**1–45.

LEE, E. Y.-H. P., and W.-H. LEE. 1986. Molecular cloning of the human *esterase D* gene, a genetic marker of retinoblastoma. Proc. Natl. Acad. Sci. USA **83:**6337–6341.

MACPHEE-QUIGLEY, K. T., T. VEDVICK, P. TAYLOR, and S. S. TAYLOR. 1986. Profile of the disulphide bonds in acetylcholinesterase. J. Biol. Chem. **260:**13565–13570.

MCTIERNAN, C., S. ADKINS, A. CHATONNET, T. A. VAUGHAN, C. F. BARTELS, M. KATT, T. I. ROSENBERRY, B. N. LA DU, and D. LOCKRIDGE. 1987. Brain cDNA clone for human cholinesterase. Proc. Natl. Acad. Sci. USA **84:**6682–6686.

MANE, S. D., C. S. TEPPER, and R. C. RICHMOND, 1983. Studies on esterase 6 in *Drosophila melanogaster*. XIII. Purification and characterization of the two major enzymes. Biochem. Genet. **21:**1019–1040.

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

MANIATIS, T., S. GOODBOURN, and J. A. FISCHER. 1987. Regulation of inducible and tissue-specific gene expression. Science **236:**1237–1244.

MYERS, M., R. C. RICHMOND, and J. G. OAKESHOTT. 1988. On the origins of esterases. Mol. Biol. Evol. **5:**113–119.

NORRANDER, J., T. KEMPE, and J. MESSING. 1983. Construction of improved M13 vectors using deoxyribonucleotide-directed mutagenesis. Gene **26:**101–106.

OAKESHOTT, J. G., C. COLLET, R. W. PHILLIS, K. M. NIELSEN, R. J. RUSSELL, G. K. CHAMBERS, V. ROSS, and R. C. RICHMOND. 1987. Molecular cloning and characterization of esterase-6, a serine hydrolase of *Drosophila*. Proc. Natl. Acad. Sci. USA **84:**3359–3363.

PEN, J., G. J. BOLKS, M. L. L. HOEKSEMA-DU PUI, and J. J. BEINTEMA. Variability in enzymic properties and structural conservation of serine esterases. (submitted).

PEN, J., A. SCHIPPER, H. A. H. RONGEN, and J. J. BEINTEMA. 1986. Differences in specificity and catalytic efficiency between allozymes of esterase-4 from *Drosophila mojavensis*. Mol. Biol. Evol. **3:**366–373.

POWELL, J. R., A. CACCONE, G. D. AMATO, and C. YOON. 1986. Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. Proc. Natl. Acad. Sci. USA **83:**9090–9093.

PRICE, N. C., and L. STEVENS. 1982. Fundamentals of enzymology. Oxford University Press, Oxford.

REED, K. C., and D. A. MANN. 1985. Rapid transfer of DNA from agarose gels to nylon membranes. Nucleic Acids Res. **13:**7207–7221.

RIGBY, P. W. J., M. DIECKMAN, C. RHODES, and P. BERG. 1977. Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113:**237–251.

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

SCHAEFFER, S. W., and C. F. AQUADRO. 1987. Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura:* evolutionary change and evidence for an ancient gene duplication. Genetics **117:**61–78.

SCHUMAKER, M., S. CAMP, Y. MAULET, M. NEWTON, K. MACPHEE-QUIGLEY, S. TAYLOR, T. FRIEDMAN, and P. TAYLOR. 1986. Primary structure of *Torpedo californica* acetylcholinesterase deduced from its cDNA sequence. Nature **319:**407–409.

SCOTT, D. 1986. Inhibition of female *Drosophila melanogaster* remating by a seminal fluid protein (ESTERASE 6). Evolution **40:**1084–1091.

SHEEHAN, K. B., R. C. RICHMOND, and B. J. COCHRANE. 1979. Studies of esterase 6 in *Drosophila*

melanogaster. III. The developmental pattern and tissue distribution. Insect Biochem. **9:** 443–450.

SIKAROV, J.-L., E. KREJCI, and J. MASSOULIE. 1987. cDNA sequences of *Torpedo marmorata* acetylcholinesterase: primary structure of the precursor of a catalytic subunit: existence of multiple 5′-untranslated regions. EMBO J. **7:**1865–1873.

SJOSTROM, M., S. WOLD, A. WEISLANDER, and L. RILFORS. 1987. Signal peptide amino acid sequences in *Escherichia coli* containing information related to final protein localization, a multivariate data analysis. EMBO J. **6:**823–831.

THROCKMORTON, L. H. 1975. The phylogeny, ecology and geography of *Drosophila.* Pp. 421–469 *in* R. C. KING, ed. Handbook of genetics. Vol. **3.** Plenum, New York.

WICKENS, M., and P. STEPHENSON. 1984. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3′ end formation. Science **226:**1045–1051.

ZOUROS, E., W. VAN DELDEN, R. ODENSE, and H. VAN DIJK. 1982. An esterase duplication in *Drosophila:* differences in expression of duplicate loci within and among related species. Biochem. Genet. **20:**929–942.

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

ROGER MILKMAN, reviewing editor

Received January 9, 1989; revision received May 15, 1989

Accepted August 23, 1989