J Mol Evol (1996) 43:241–252

JOURNAL OF MOLECULAR EVOLUTION

Duplication and Divergence of the Genes of the α-Esterase Cluster of *Drosophila melanogaster*

Charles Robin,[1,2] Robyn J. Russell,[1] Kerrie M. Medveczky,[1] John G. Oakeshott[1]

[1] CSIRO Division of Entomology, GPO Box 1700, Canberra ACT 2601, Australia
[2] Division of Botany and Zoology, Australian National University, Canberra ACT 0200, Australia

Received: 18 January 1996 / Accepted: 12 March 1996

Abstract. The α-esterase cluster of *D. melanogaster* contains 11 esterase genes dispersed over 60 kb. Embedded in the cluster are two unrelated open reading frames that have sequence similarity with genes encoding ubiquitin-conjugating enzyme and tropomyosin. The esterase amino acid sequences show 37–66% identity with one another and all but one have all the motifs characteristic of functional members of the carboxyl/cholinesterase multigene family. The exception has several frameshift mutations and appears to be a pseudogene. Patterns of amino acid differences among cluster members in relation to generic models of carboxyl/cholinesterase protein structure are broadly similar to those among other carboxyl/cholinesterases sequenced to date. However the α-esterases differ from most other members of the family in: their lack of a signal peptide; the lack of conservation in cysteines involved in disulfide bridges; and in four indels, two of which occur in or adjacent to regions that align with proposed substrate-binding sites of other carboxyl/cholinesterases. Phylogenetic analyses clearly identify three simple gene duplication events within the cluster. The most recent event involved the pseudogene which is located in an intron of another esterase gene. However, relative rate tests suggest that the pseudogene remained functional after the duplication event and has become inactive relatively recently. The distribution of indels also suggests a deeper node in the gene phylogeny that separates six genes at the two ends of the cluster from a block of five in the middle.

Key words: Esterase — Gene cluster — Pseudogene — *Drosophila melanogaster*

Introduction

The α-esterases of *D. melanogaster* are classified within a superfamily of largely hydrolytic enzymes that share a common structure called the α/β hydrolase fold (Ollis et al. 1992). Although members of this superfamily do not necessarily share significant sequence similarity, they have a similar tertiary structure and a common secondary-structure topology, and the positions of three ‘‘catalytic’’ residues on this topology are conserved (Ollis et al. 1992).

Two of the proteins whose empirically derived structures were used to define the α/β hydrolase fold, a lipase from a yeast (*Geotrichum candidum,* GcLip) and acetylcholinesterase from an eel (*Torpedo californica,* TcAChE), belong to the carboxyl/cholinesterase multigene family (E.C.3.1.1; Oakeshott et al. 1993). Membership in this family is determined entirely by sequence similarity. In the case of TcAChE and GcLip the sequences are 24% identical and 52% similar but superimposition of their tertiary structures shows that 399 of the 540 α carbons (74%) overlap with an r.m.s. deviation of 1.90 Å (Ollis et al. 1992). The general tertiary structures of other proteins in the carboxyl/cholinesterase multigene family can therefore be predicted with some confidence (Cygler et al. 1993; Lotti et al. 1993; Oakeshott et al. 1993; Jones et al. 1994). In a study by Cygler et al. (1993) 32 carboxyl/cholinesterase sequences were com-

Correspondence to: C. Robin

pared with respect to the structures of TcAChE and GcLip. The residues conserved throughout the family were shown to occur in the catalytic triad, or on the edges of the major β sheet, or in the hydrophobic core, or were involved in either disulfide or salt bridges.

While the general structure and catalytic machinery of carboxyl/cholinesterases are essentially conserved, their diverse substrate specificities appear to be determined by variation in residues that line the active site gorge or that occur on loops on the lip of the gorge. For example, the active site gorge of TcAChE is lined with aromatic residues that apparently guide acetylcholine to the active site (Sussman et al. 1991; Ripoll et al. 1993), while some lipases have a lid covering the active site gorge that is thought to open when the enzyme is at an oil/water interface (Shrag and Cygler 1993). Mutagenesis studies of acetylcholinesterase have shown that two amino acid changes in the ''acyl pocket'' of the active site gorge are enough to confer butyrylcholinesterase activity while reducing acetylcholinesterase activity (Harel et al. 1992; Vellom et al. 1993). In some insect species amino acid replacements at three sites have also been shown to protect the enzyme from inhibition by organophosphate (OP) insecticides and thus confer ''target site'' resistance (Mutero et al. 1994).

In this paper we analyze the sequences of the 11 carboxyl/cholinesterase genes that form the α-esterase cluster of *D. melanogaster* (Russell et al. 1996). The α-cluster is located within cytological region 84D3-E2 and is so far known to encode four biochemically defined esterase phenotypes: EST9, EST23, MCE, and ALI (Spackman et al. 1994). EST9 (formerly EST C; Wright and MacIntyre 1963; Healy et al. 1991) is the major α-naphthyl-acetate-hydrolyzing isozyme of *D. melanogaster* and appears to be conserved throughout the *Drosophila* radiation (Gillespie and Kojima 1968; Morton and Singh 1985). The EST23 isozyme has an ortholog (E3) in the sheep blowfly, *Lucilia cuprina* (Spackman et al. 1994), that has been implicated in metabolic resistance to OP insecticides (Hughes and Raftos 1985). No isozyme phenotypes have been identified for MCE and ALI, which are defined instead by radiometric and spectrophotometric assays for malathion and methyl butyrate hydrolysis, respectively (Spackman et al. 1994). An MCE activity which maps very closely to E3 in *L. cuprina* (Smyth et al. 1994) and an ALI activity mapping to the homologous chromosome in the house fly *Musca domestica* (Weller and Foster 1993) are also associated with metabolic OP resistance.

The phylogenetic distribution of the enzymes suggests that at least one esterase of the α-cluster may be traced back at least 80 million years to the divergence of the drosophilid and calliphorid lineages (Beverley and Wilson 1984). It further suggests a predisposition for the α-esterase genes to mutate to confer metabolic resistance to OPs. We present phylogenetic and functional analyses of the α-cluster sequences of *D. melanogaster* in order to reconstruct the phylogeny of the cluster and identify sequence characteristics that might underlie distinctive biochemical functions of the α-esterases.

Materials and Methods

All DNA fragments analysed in this paper were derived from the overlapping bacteriophage clones λPK-9, -14, -18, -19, -26, -27, -28, -30 described in Russell et al. (1996). Restriction fragments that Russell et al. (1996) found to hybridize to consensus esterase oligonucleotides were subcloned into pBCKS$^+$ (Stratagene). The Erasabase system (Promega) was used to create nested deletions of these subclones that were then sequenced using dye-labeled primers (ABI M13 forward and reverse primers) on an automated sequencer (Applied Biosystems model 370A). In total 35 kb was sequenced on both strands and a further 7 kb on one strand only (Fig. 1).

The entire coding regions of the 11 esterase genes were sequenced in both strands with the exception of the 5′ end of α*E8,* which has yet to be identified. The large (2,742 nt) first intron of α*E10* was sequenced completely in only one direction and about 100 nt of the first intron of α*E7* was not sequenced in either strand. Some sequence of regions adjacent to esterase ORFs was also obtained. Nonesterase open reading frames (ORFs) were compared to the nonredundant databases using the NCBI BLAST email server (Altschul et al. 1990).

Initially the coding regions of the α-esterases were translated and aligned with each other using the default parameters (gap weight = 3.0, gap length weight = 0.1) and default scoring matrix of the GCG global alignment package ''pileup'' (Genetics Computer Group 1994). For this purpose the pseudogene (α*E*Ψ) was kept in frame by inserting x's into the nucleotide sequence before translation. Regions of unstable alignment were then identified by altering the gap weight to 5.0 or 10.0, aligning only a subset of six sequences, and comparing the amino acid alignment with a pileup alignment of the corresponding nucleotide sequences (default parameters; gap weight = 5.0, gap length weight = 0.3). Three regions, up to codon 4, between codons 345 and 395, and between 502 and 519 (numbering here and below is according to α*E1* unless otherwise stated), were found to be sensitive to these various alignment regimes and were not used in tree building.

Alignment of the α-esterase amino acid sequences with other esterases was based upon the alignment of Cygler et al. (1993). Indels (insertions or deletions) were placed outside known secondary structural elements of other esterases.

The Phylip (Felsenstein 1991) and PAUP (Swofford 1993) packages were used to construct unrooted phylogenetic trees. In the former, neighbor-joining algorithms were used with a PAM amino acid transition matrix (programs ''protdist'' and ''neighbor''). The distance tree was then drawn with ''drawtree.'' Paup version 3.1 was used to construct parsimony trees. A series of transition matrices was used in the parsimony analyses. Matrices weighting transitions and transversions differently were used for the nucleotide data, whereas ''step matrices'' scoring the minimum number of nonsynonymous changes between different codons were used for amino acid comparisons (''protpars'' matrix; Swofford and Begle 1993). Bootstrap values were based upon 100 randomly resampled replicates. The proportion of synonymous differences was calculated with all gap sites removed using MEGA (Kumar et al. 1993).

Results

Organization of the Cluster

We have sequenced about 42 kb in five segments spanning 60 kb (Fig. 1). The segments sequenced include the coding regions of 11 esterase genes (1,614–

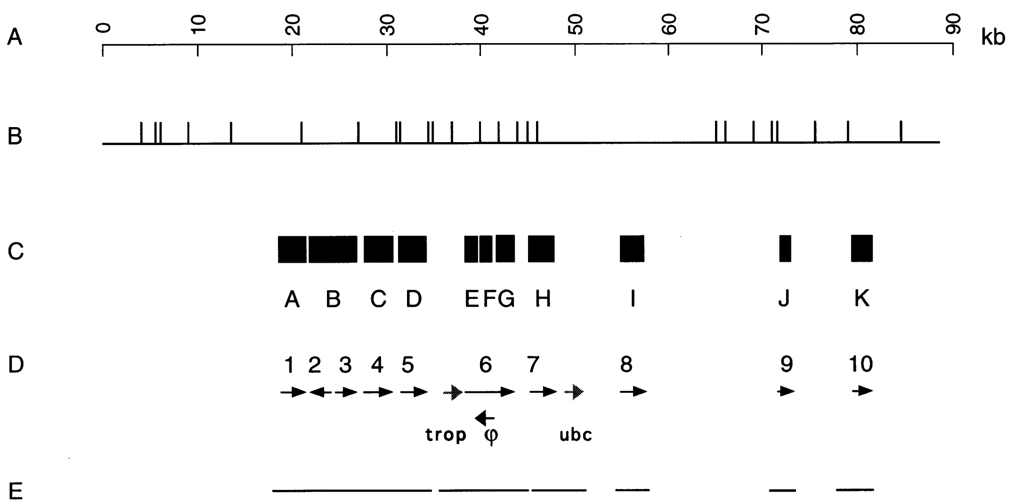


Fig. 1. The genomic organization of the α-esterase cluster. **A** Scale in kb. **B** Map of *Eco*RI sites from Russell et al. (1996). **C** Restriction fragments that hybridize to esterase consensus oligonucleotides from Russell et al. (1996). **D** Orientation and positioning of esterase genes α*E1*–α*E10,* the pseudogene (α*E*Ψ) which occurs in intron II of α*E*6, and the tropomyosin-like (*trop*) and ubiquitin-conjugating (*ubc*) ORFs. **E** The regions sequenced are represented by *lines.*

1,716 bp each) and two unrelated ORFs that are similar to a ‘‘tropomyosin-related’’ protein (gp|L10335|HUMNSPC_1) and a ubiquitin-conjugating enzyme (UBC; gp|U00035|CELR01H2_6) (40/98 and 62/122 amino acid identities, respectively). The UBC-like sequence is 164 amino acids long and contains several conserved motifs including a cysteine that is required for ubiquitin-thiolester formation (Jentsch et al. 1990). The sequence of the tropomyosin-like ORF is incomplete but there are two separate regions of similarity (20/50 and 20/48 amino acid identities) between it and the most similar tropomyosin-like protein in the database. The argument for homology of this ORF with a tropomyosin-related protein is strengthened by the conservation of the distance between the two motifs (42–43 amino acids). The sequences of the UBC and tropomyosin-like ORFs and the sequences of the α-esterases can be obtained from Genbank (accession numbers U51043–U51054).

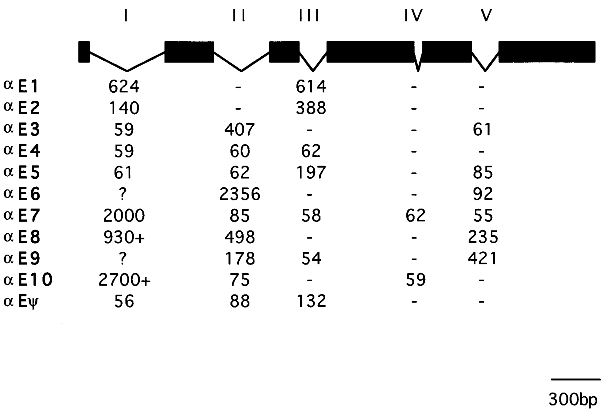
All but two of the α-esterase genes (α*E2,* α*E*Ψ) are oriented in the same direction. The spacing between the functional genes ranges from 0.5 to 9 kb, with the genes of the ‘‘left side’’ (eight in 30 kb; Fig. 1) more tightly clustered than those of the ‘‘right’’ (four in 35 kb). One of the genes, α*E*Ψ, appears to be a pseudogene and is located within the intron of another gene (α*E*6). Its coding region has a 7-nucleotide insertion, a 17-nucleotide deletion, and a one nucleotide insertion with respect to all the other genes. Its first in-frame stop codon occurs 275 nucleotides from its start methionine. It also has the lowest third-position GC content (50.7% compared to 55.3–72.7%) and highest third-position T content (29.9% compared to 15.3–23.6%) of all the α-esterase genes. The deduced amino acid sequence of αEΨ also has a few unusual features; most notably it has a tyrosine in the place of the catalytic triad histidine.

α-*Esterase Gene Structure*

Multiple alignment of the nucleotide sequences reveals five sites within a gene where introns may be found (designated I–V; Fig. 2). All the genes, including α*E*Ψ, have introns in at least two of these sites and α*E7* has them in all five. The splice and acceptor sequences that flank the introns match the consensus of 209 *D. melanogaster* introns examined by Mount et al. (1992). The α-esterase consensus can be summarized as $A_{69}G_{56}G_{100}T_{100}A_{55}A_{81}G_{84}T_{68}$ for the splice donor sequences and $C_{76}A_{97}G_{100}$ for the splice acceptor sequences (subscripts are the percent occurrences). There is only one exception to the GT/AG splice site rule and that occurs in intron II of α*E*Ψ, where AG is replaced by TG. This is interpreted as further evidence that α*E*Ψ is a pseudogene.

A start methionine can be identified for eight of the genes. It lies between three and 18 codons 5′ of intron I, depending on the gene. The codons for these start methionines all have a purine at position −3 and generally conform to the loose consensus described by Cavener and Ray (1991). We are uncertain of the identity of the start methionines in α*E*6 and α*E*9 and whether their genes contain the first intron. They both have an in-frame potential start methionine; however, these may be spliced out since they also have splice acceptor sites that align with the splice acceptor sites of the other α-esterase genes. The start methionine for α*E*8 is also undetermined, whereas the identities of the start methionines and the presence of introns in α*E5,* α*E7,* and α*E10* have been confirmed by cDNA analyses (data not shown).

Eight of the 11 α-esterase genes have at least one consensus polyadenylation signal (AATAAA) and the other three (α*E7,* α*E8,* and α*E9*) have potential polyad-



	I	II	III	IV	V
αE1	624	-	614	-	-
αE2	140	-	388	-	-
αE3	59	407	-	-	61
αE4	59	60	62	-	-
αE5	61	62	197	-	85
αE6	?	2356	-	-	92
αE7	2000	85	58	62	55
αE8	930+	498	-	-	235
αE9	?	178	54	-	421
αE10	2700+	75	-	59	-
αEψ	56	88	132	-	-

Fig. 2. Distribution of introns across the five sites in the 11 α-esterase genes. The size of introns is given in bp, and the absence of an intron is represented by a *dash.* Sites where it is unclear if an intron exists are represented by a *question mark.* Exact coordinates for the five sites are given in Fig. 4.

enylation signals that vary from the consensus by one nucleotide (AACAAA, AATATA, CATAAA, respectively). The polyadenylation signals lie between 10 and 499 bp 3′ of the respective stop codons. The program Signal Scan (Prestridge 1991) was used to search for promoter elements. None were found to be shared by all genes of the cluster. Seven contain 5′ TATA boxes within 800 bp of the start codon (α*E1,2,3,5,7,9,10*) and the most common other 5′ motifs found were GC boxes (α*E1,9*), GAGA boxes (α*E3,7,10*), and zeste elements (α*E1,2,3,6,7,10*).

α-*Esterase Protein Structure*

We have used the weighted matrix method of von Heijne (1986; as implemented by the program sigseq) to predict that none of the α-esterases have N-terminal signal peptides (scores range from −2 to −12.2, which is well within the range of cytosolic sequences). Consequently we expect that the nascent α-esterase polypeptides are localized in the cytosol rather than the endoplasmic reticulum and are therefore not subject to N-linked glycosylation. Although this does not mean that the esterases remain in the cytosol, it does limit their intracellular fate. Since EST9 segregates with the soluble fraction in density centrifugations (Spackman et al. 1994) it probably remains in the cytosol. EST23, on the other hand, segregates with the microsomal fraction centrifugations (Spackman et al. 1994) and thus is probably attached or moved across a membrane of a small organelle such as a peroxisome.

Pairwise amino acid identities among the α-esterases range from 37 to 66% (average 56%). The distribution of variation across the molecule is remarkably similar to that among other carboxyl/cholinesterases (Fig. 3). In general, the N-terminal subdomain (defined by Cygler et al. 1993 as the first 322 amino acids of TcAChE) and the residues in the β strands are more conserved and align well with TcAChE, while the C-terminal subdomain contains two of the three stretches of uncertain alignment (345–395, 502–519: numbered according to αE1; Fig. 4), both of which align with residues located on the surface of TcAChE.

Some of the sequences that are most conserved among the α-esterases correspond to features that are highly conserved among other carboxyl/cholinesterases and are essential for their catalytic function. These sequences include residues of the catalytic triad (S207, E340, and H459) and the regions surrounding them, specifically, the ‘‘nucleophilic elbow’’ ($T_{100}XF_{82}G_{100}E_{73}S_{100}A_{82}G_{100}$; 202–209; subscripts indicate the percentage of the 11 esterases that have the conserved amino acid), ‘‘acid turn’’ ($S_{91}F_{64}E_{100}G_{100}$; 338–341) and ‘‘histidine loop’’ ($G_{100}V_{82}XH_{91}A_{55}D_{100}D_{73}L_{91}S_{64}Y_{100}$; 456–465) and the ‘‘oxyanion hole’’ ($G_{100}G_{100}G_{64}$; 124–126; Fig. 4). This conservation suggests that all the α-esterases (except αEΨ) may be catalytically active.

Most of the other regions that are highly conserved in the α-esterases align with the motifs of other carboxylcholinesterases (Fig. 3) and are likely to maintain the generic esterase tertiary structure. These include tight turns, which often have irregular bonding angles, and amino acids that maintain the hydrophobic core (Cygler et al. 1993). The most highly conserved motif ($G_{100}S_{82}E_{100}D_{100}C_{100}L_{100}Y_{91}L_{91}N_{100}V_{82}Y_{82}$; 96–106) contains both a cysteine residue involved in a putative disulfide bridge (77–100) and a glutamic acid (98) involved in a putative salt bridge with residue 61.

The most obvious difference between the similarity plot comparing the α-esterases with one another (Fig. 3A) and the plot comparing other esterases (Fig. 3B) is a conserved motif in the former ($P_{91}T_{91}Y_{91}L_{73}Y_{100}R_{100}F_{91}D_{100}F_{82}D_{100}S_{100}$; 429–439) at and beyond the sequence that aligns with strand β8 from the major β sheet. The α-esterases have seven to eight more amino acids than TcAChE at this position (Fig. 4) and secondary-structure predictions using the PHD secondary-structure prediction program (Rost and Sander 1993) suggest that they contribute to a loop rather than extend the β strand.

There are three other major indel differences between the α-esterases and TcAChE (excluding those in the three regions of uncertain alignment). Examination of the TcAChE model (by the Rasmac graphics program; Sayle 1994) shows that all three indels occur on the outside of the molecule and two are located around the lip of the active site groups. The largest of these three indels is an 11-amino-acid deletion in the α-esterases (between residues 93 and 94). The corresponding region of TcAChE contains a small α helix ($\alpha_{b3,2}$) on a loop stabilized by a flanking disulfide bridge. The cysteines that form the disulfide bridge are conserved among 28 of the 29 catalytically active carboxyl/cholinesterases analyzed by Cygler et al. (1993) and the residues between them, which are missing in the α-esterases, are thought to have a role

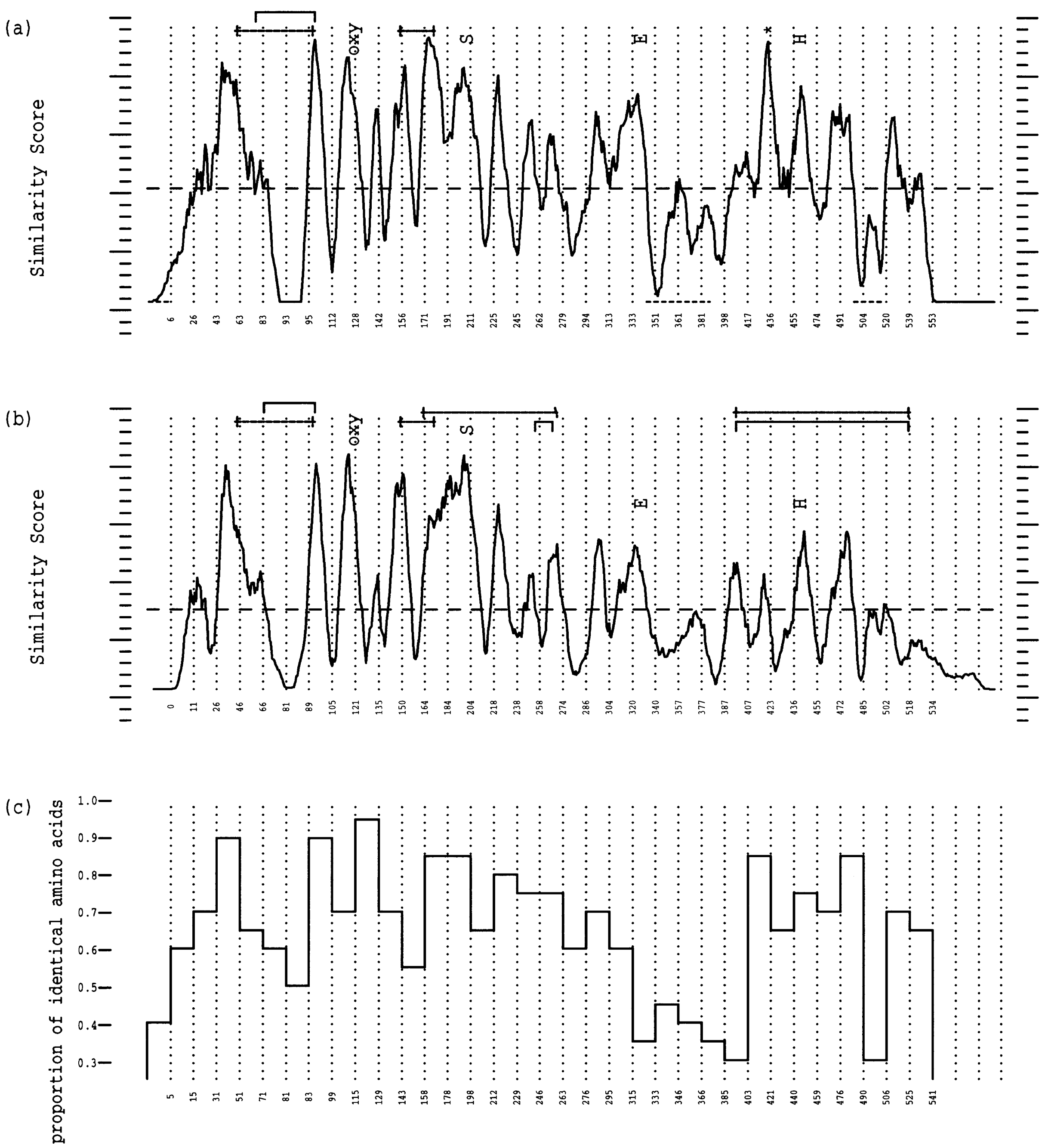


Fig. 3. Distribution of variation across esterase primary structure. **A** A similarity plot showing the distribution of variation across the *D. melanogaster* α-esterases (excluding αEΨ), constructed using the GCG ''plotsimilarity'' program. Peaks represent conserved motifs and troughs represent regions of high variability. The positions of the catalytic triad residues are indicated as *S, E,* and *H* and the oxyanion hole motif is indicated as *oxy.* Putative disulfide bridges are marked as ┌──┐ and putative salt bridges as ⊢──⊣. The *asterisk* indicates the most distinctive motif of the α-esterases (see text), and the regions of uncertain alignment (see Methods) are indicated at the bottom of the plot as - - - - - -. The x-axis is *numbered* according to αE1 residues. **B** A similarity plot of 29 members of the carboxyl/cholinesterase multigene family analyzed by Cygler et al. (1993) calculated as in **A** and numbered according to TcAChE. **C** A histogram representing the proportion of identical residues in αE4 and αEΨ. The sequences were divided into bins representing 20 amino acid positions of the alignment used in **A.** The x-axis is numbered according to αE4.

in substrate binding. In addition to the indel difference the putative disulfide bridge in this region also seems to be slightly altered, since the first cysteine occurs seven amino acids before the corresponding cysteines in other members of the carboxyl/cholinesterase multigene family. The other two large indel differences between the α-esterases and TcAChE are located at 288–290 and 313–317 in the αE1 sequence. The 288–290 indel aligns with or next to the ''peripheral binding sites'' of TcAChE (Cygler et al. 1993). Adjacent to these are aromatic residues that are conserved among vertebrate cholinesterases and form part of the aromatic lining of the

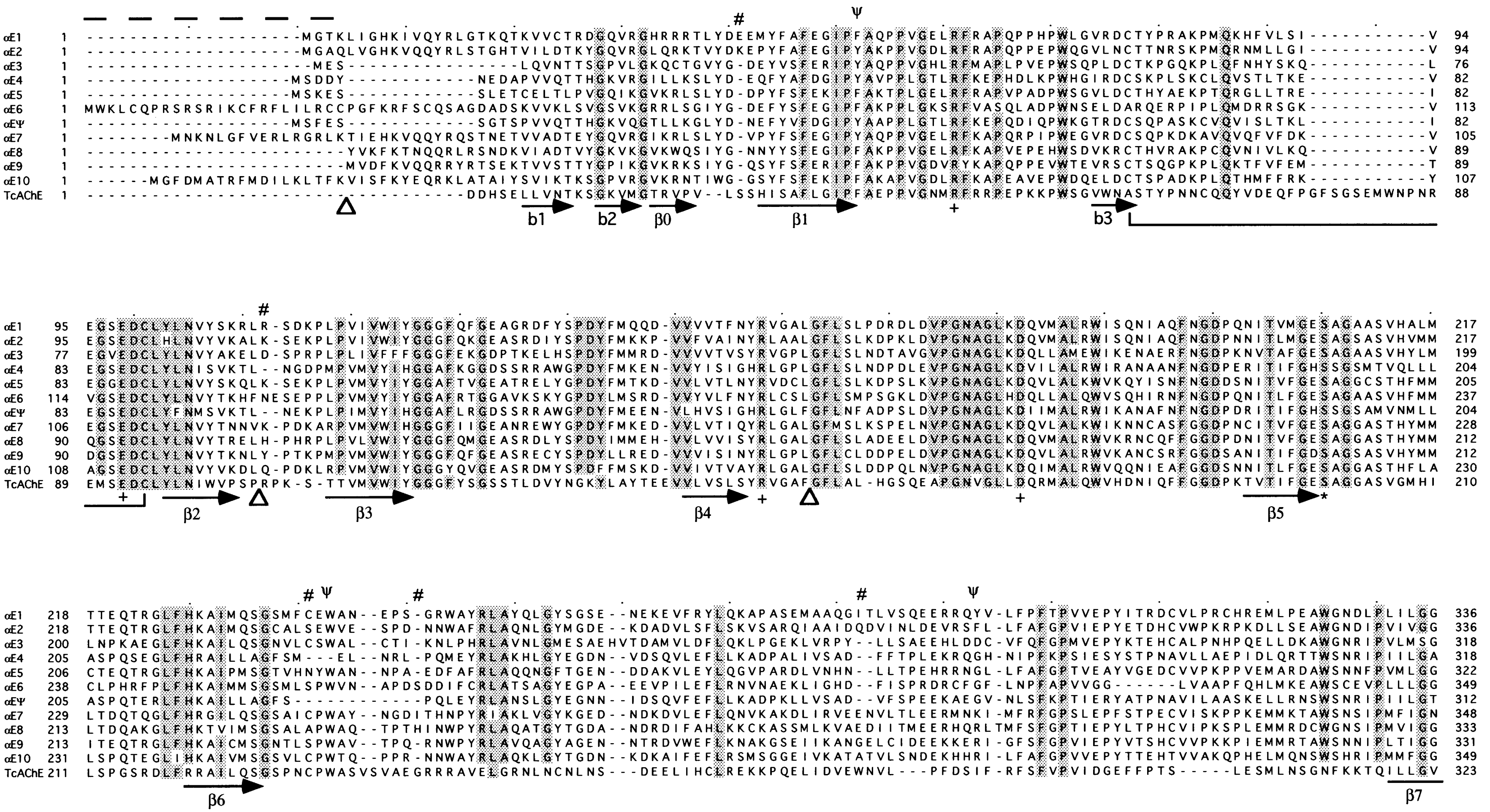


Fig. 4. Alignment of the α-esterase amino acid sequences with the sequence of TcAChE (Shumacher et al. 1986). Regions of uncertain alignment (see Methods) are *overlined with a dotted line.* Residues identical in at least 11 sequences are *shaded.* The residues in TcAChE that are part of β strands are underlined by *arrows.* The single putative disulfide bridge that the α-esterases share with other members of the carboxyl/cholinesterase multigene family is represented by |_____|. Residues that probably contribute to conserved salt bridges are indicated with a *plus sign* and those that form the catalytic triad by *asterisk.* Phylogenetically informative indels (see Fig. 5B) are marked by *#*. Sites where αEΨ changes reading frame are indicated by Ψ and *triangles* represent the positioning of introns.

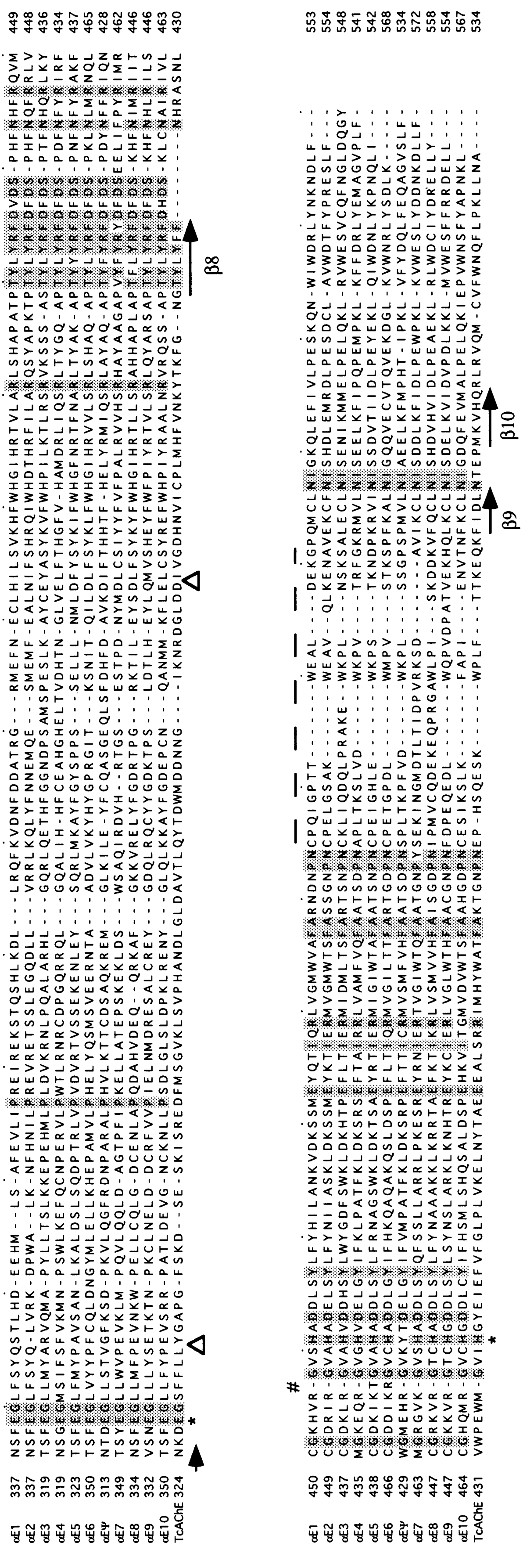


Fig. 4. Continued.

active site gorge (Sussman et al. 1991). The 313–317 indel is five amino acids long and occurs at least 9 Å away from the edge of the active site gorge.

Phylogenetic Analyses

Parsimony and distance analyses were used to reconstruct a phylogeny of the *D. melanogaster* α-esterases. Trees with the same significant nodes were obtained with both methods using either nucleotide or amino acid data. The neighbor-joining tree in Fig. 5A also includes esterases associated with OP resistance in the mosquitoes *Culex pipiens* and *C. quinquefasciatus* (CpESTB1; Mouchès et al. 1990 and CqESTA2, also called Estα2; Vaughan and Hemingway 1995) and the aphid *Myzus persicae* (MpE4; Field et al. 1993) as well as two esterases encoded by the β cluster of *D. melanogaster* (DmEST6 and DmESTP; Collet et al. 1990). Although neither of the two *Culex* esterases is orthologous to any of the *D. melanogaster* α-esterases, they cluster more closely to them than to the aphid esterase or the *D. melanogaster* β-esterases.

Two nodes within the *D. melanogaster* α-esterase cluster have bootstrap scores of 100. These link αE4 to αEΨ and αE1 to αE2. Both of these nodes involve genes oriented in opposite directions although, as we discuss below, α*E1* and α*E2* are physically adjacent genes whereas α*E4* and α*E*Ψ are not. The only other significant node within the cluster is between the equivalently oriented and adjacent α*E8* and α*E9* genes (bootstrap score of 87).

Sliding window analyses (window size of 100 codons) showed no obvious spatial heterogeneity along the primary sequence in the proportion of silent site differences in any pairwise comparison of the α-esterase genes (not shown). This suggests that there has been no recent partial gene conversion nor nonreciprocal recombination events among the α-cluster members. The overall proportion of pairwise silent site differences (p) ranges from 0.60 to 0.80. Thus even the values for the α*E4*/α*E*Ψ (p = 0.64) and α*E1*/α*E2* (p = 0.72) comparisons approach saturation. Orthologous comparisons of other *Drosophila* genes show that silent site saturation is not achieved in comparisons of *D. melanogaster* and *D. pseudoobscura* (p between 0.24 and 0.48 in six genes compared by Riley 1989) but is approached in comparisons between *D. melanogaster* and *D. willistoni* (p = 0.61 and 0.64 for *Sod* and *Adh,* respectively). Various lines of evidence put the *D. melanogaster/D. pseudoobscura* and *D. melanogaster/D. willistoni* splits at 25–40 and 30–40 mya, respectively (Powell and DeSalle 1995). We therefore propose that even the most recent gene duplication events happened at least 25 million years ago and possibly even before the separation of the willistoni and melanogaster lineages.

Since insertions and deletions probably occur less fre-

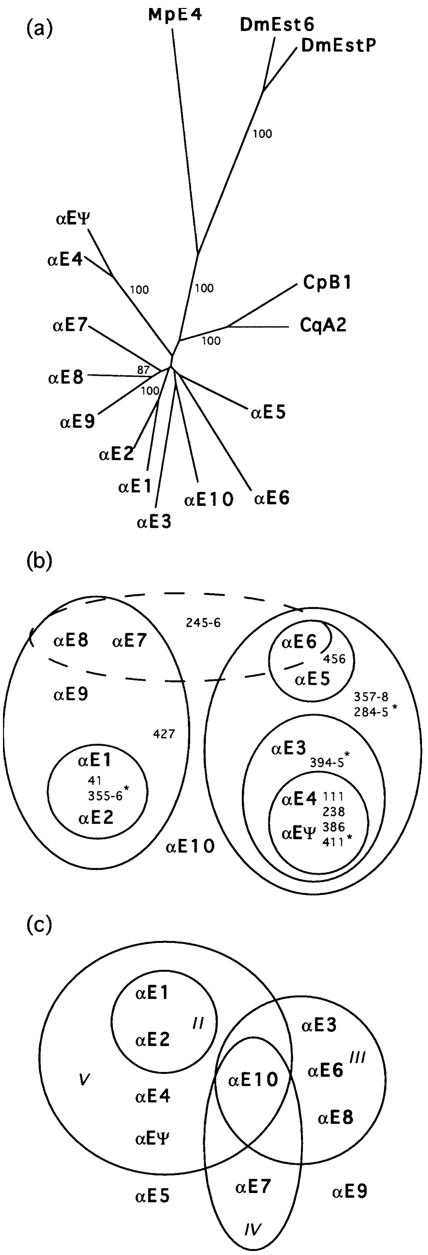


Fig. 5. **A** An unrooted neighbor-joining tree of the α-esterase amino acid sequences, EST6 and ESTP from *D. melanogaster* (Collet et al. 1990), ESTB1 from *Culex pipiens* (CpESTB1; Mouchès et al. 1990), ESTA2 from *C. quinquefasciatus* (CqESTA2; Vaughan and Hemingway 1995), and esterase E4 from *M. persicae* (MpE4; Field et al. 1993). *Numbers* on the branches indicate the percentage of bootstrapped replicates that group the taxa together when the data set is randomly resampled 100 times. The branch lengths are proportional to distance. **B** A Venn diagram partitioning the α-esterases by ''phylogenetically informative'' indels (i.e., indels which divide the taxa into two groups each containing more than one taxon; this excludes sites where there are indels of multiple lengths among the taxa). Indels are *numbered* according to their position on the αE1 amino acid sequence and those that occur in the regions of uncertain alignment are marked with an *asterisk.* The *dotted circle* highlights the only intersecting circle, and thus the only inconsistency in the indel data. **C** A Venn diagram partitioning the α-esterases according to the presence/absence of introns at sites II to V (see Fig. 2).

quently than either amino acid or nucleotide substitutions, they should retain phylogenetic signal for a longer period of evolutionary time. In the α-esterases there are 12 phylogenetically informative indels (Fig. 4). A Venn diagram relating the 11 α-esterase genes on the basis of these indels produces only one intersection indicative of phylogenetic inconsistency (Fig. 5B). This is due to indel 245–246, which groups of αE6 with αE7 and αE8, and conflicts with four other indels (284–285, 357–358, 427, 455–456) that separate αE6 from αE7 and αE8. Otherwise the Venn diagram produces a similar set of phylogenetic relationships to those evident from the sequence data above. In particular it supports both the αE1/αE2 and αE4/αEΨ groupings and, except for the aberrant indel 245–246, the αE8/αE9 grouping as well. It also suggests another deeper node separating five contiguous genes in the middle of the cluster (α*E3,4,5,6,Ψ*) from the six at the two ends (α*E1,2* on the left and α*E7,8,9,10* on the right). The aberrant distribution of indel 245–246 requires that there were either independent insertion events or possibly ancient gene conversion events.

Another set of characters that could be used to construct a phylogeny is the presence/absence of introns at the five intron sites found among the α-esterase genes (Fig. 2). Intron I is not phylogenetically informative but a Venn diagram using the other four sites produces two intersections (Fig. 5C). The αE1/αE2 and αE4/αEΨ groupings are still observed but αE8 and αE9 differ with respect to intron III and there is no clear distinction between the αE1,2,7,8,9,10 and αE3,4,5,6,Ψ groups. The observed intron distribution can only be achieved by multiple independent gain/loss events. As few as six gain/loss events are required under the most parsimonious model and eight if the introns are forced onto a tree that splits the taxa into a middle group and a flanking group as do the indel data (data not shown).

Discussion

With 11 esterase genes in 60 kb, the α-esterase cluster of *D. melanogaster* is one of the largest enzyme-encoding clusters so far described. The genes are irregularly spaced, with twice as many in the left 30 kb (α*E1–6*) as in the right 30 kb (α*E7–10*). All but two of the genes occur in the same orientation, and sequence similarity within the cluster is limited to the coding regions. Two nonesterase open reading frames have also been found within the cluster and these have similarity to genes encoding a ubiquitin-conjugating enzyme and tropomyosin. These may provide useful markers for gene duplications if other copies or their remnants are found in the as-yet-unsequenced portions of the cluster. They may also help establish orthology with esterases in distantly related species.

A remarkable feature of the cluster's organization is the presence of a pseudogene (α*EΨ*) within an intron of another gene. The pseudogene has at least five disabling mutations, including three frame shifts, a mutated intron splice acceptor, and an incomplete active site triad. It does not appear to be a processed pseudogene since it contains the remnants of three introns and does not possess a poly-A tail. Nevertheless, there are 332 nucleotides at its 5′ end and 128 nucleotides at its 3′ end which separate α*EΨ* from the intron boundaries of α*E6*, indicating that regulatory elements may well have been duplicated as well as the coding region of the gene. This is consistent with evidence covered below that α*EΨ* has not always been a pseudogene.

The esterases of the α-cluster have diverged substantially from one another (amino acid similarity ranging from 37 to 66%). However, the distribution of variation corresponds well with the variation observed among other active carboxyl/cholinesterases. Many of the conserved motifs correspond to features that maintain the α/β hydrolase fold of other carboxyl/cholinesterases, such as crucial turns and salt and cysteine bridges. Other conserved residues include the active site triad, the oxyanion hole, and the nucleophilic elbow, which are crucial for catalysis, suggesting that all the α-esterases (except αEΨ) may be catalytically active.

The α-esterases differ from most other carboxyl/cholinesterases by four large indels. Two of these are located around the rim of the active site gorge and thus probably have a role in substrate specificity. A third is a seven-amino-acid insertion, which is distant from the residues involved in substrate guidance and catalysis in the model of TcAChE, and its function is unclear.

All four of these indels occur in only two other carboxyl/cholinesterase sequences published to date: the ESTA2 and ESTB1 esterases from *Culex* mosquitoes (Mouchès et al. 1990; Vaughan and Hemingway 1995). The *Culex* esterases also lack a signal peptide and have an unusual disulfide bridge structure; and the CpESTB1 gene shares intron positions II and V with the *D. melanogaster* α-esterase genes (the intron structure for the gene encoding CqESTA2 is unknown). Furthermore, the *Culex* genes encoding the A- and B-type esterases are closely linked genetically (Villani et al. 1983; Wirth et al. 1990). Taken together these data support the proposition of Russell et al. (1996) that the *Culex* esterases may represent the homologous cluster to the *D. melanogaster* α-esterase cluster. This is particularly interesting since the A and B esterases, like at least one of the α-esterases of *L. cuprina*, can mutate to confer OP resistance (Wirth et al. 1990).

We have addressed the phylogenetic relationship among the α-esterase genes of *D. melanogaster* using three character sets: sequences, indels, and introns. The sequence analysis was based on regions of unambiguous alignment but the high level of variation nevertheless restricted resolving power to terminal nodes. The indel data were internally consistent and provided good, albeit few, characters. The value of the intron data, on the other hand, was reduced by the need to invoke multiple independent gain or loss events to explain their distribution.

All three data sets support nodes linking αE1 to αE2 and αE4 to αEΨ. The sequence and indel data also support the αE8/αE9 node. The presence of these three nodes suggests that the recent evolution of the cluster has been by progressive addition of genes through single gene duplication events.

Two of these recent gene duplication events have involved genes oriented in opposite directions (α*E1*/α*E2*, α*E4*/α*E*Ψ) while the third resulted in α*E8* and α*E9* positioned head to tail. Formally we do not know which of the duplicate genes is ancestral in any of the three gene duplications. However, a notable feature of the cluster is that nine of the 11 genes are oriented in the same direction, and the two genes facing in the other direction, α*E2* and α*E*Ψ, may well be the most recent additions to the cluster. Graham (1995) suggested that groups of genes oriented in the same direction are more likely to undergo unequal recombination than groups of genes oriented in opposing directions, and they are therefore likely to be more homogenized. If the most recent duplication events of the α-cluster resulted in insertion of inverted copies in what was once a tandem array it would subsequently restrict the opportunity for homogenization by unequal recombination. This is at least consistent with our failure to detect recent intergenic recombination events.

However, the structure in the indel data suggests that homogenization might have been restricted even before the two recent duplications. Furthermore, if unequal recombination were a major force in the evolution of the cluster, one might expect genes physically close to be more closely related. Instead the indel data suggest that α*E1* and α*E2* are more closely related to more distant genes of the cluster (α*E7,8,9*).

The failure to resolve this or any other nonterminal nodes in the phylogeny using sequence data probably reflects the age of the genes of the cluster. Evidence from orthologous gene comparisons of other genes across *Drosophila* species indicate that silent site saturation is generally achieved in comparisons between *D. melanogaster* and *D. willistoni.* The divergence between these has been placed at 30–40 mya, based on biogeographical calculations and the fossil record (Powell and DeSalle 1995). Since the silent site differences between all genes of the cluster including α*E4*/α*E*Ψ and α*E1*/α*E2* have approached saturation, the genes of the α-cluster must have been present through most of the radiation of the sophophoran subgenus. This is supported by biochemical evidence that three α-esterase phenotypes associated with the cluster (EST23, ALI, and MCE) have apparent homologs in *L. cuprina* (Spackman et al. 1994), the ancestors of which diverged from the *Drosophila* lineage about 80–100 mya (Beverley and Wilson 1984).

It will be particularly interesting to examine α*E*Ψ from other species of *Drosophila*, not only because it stems from the most recent duplication, and its absence in some species may place a date on the duplication event, but also because the duplication event and the inactivation event may not have occurred at the same time. The calculations above suggest that the duplication happened before the divergence of the obscura and melanogaster groups, and we are confident that α*E*Ψ is now inactive in *D. melanogaster* since it has three frame-shift mutations and a point mutation in an intron splice acceptor sequence. However, two observations suggest α*E*Ψ has not always been a pseudogene. Firstly, the distribution of the changes down the α*E*Ψ branch of the phylogeny is not random across the molecule, as would be expected if it had always been a pseudogene. All but three of the 79 residues that are invariant amongst the active genes are also conserved in the pseudogene, and furthermore most of the differences between αE4 and αEΨ occur at sites located in troughs on the similarity plot comparing all the α-esterases (Fig. 3C). Secondly, the relatively similar branch lengths leading to αE4 and αEΨ (Fig. 5A) suggest that the pseudogene and the functional α*E4* gene have been changing at about the same rate, which would not be expected if α*E*Ψ had been nonfunctional since the time of duplication. In support of this, application of Tajima's 1D relative rate test (Tajima 1993) to the amino acid sequences of αE4 and αEΨ reveals no significant differences in rates down the two lineages ($\chi^2 = 0.2$; 1 *df* $= 3.84$; $P < 0.05$). If we assume α*E*Ψ had been inactive since the time of duplication we would expect more nonsynonymous changes down the αEΨ branch, especially since the proportion of synonymous differences between α*E*4 and α*E*Ψ approaches saturation (0.64). An approximate simulation in which we reconstructed the ancestral sequence of αE4/αEΨ using αE5 as an outgroup, and in which we assumed that nonsynonymous sites in the pseudogene changed at the same rate as synonymous sites, suggests we would expect five times as many amino acid changes down the αEΨ lineage as are observed (118 rather than 24—this excludes the regions of uncertain alignment and sites at which αE4, αEΨ, and αE5 are all different from each other).

Thus the patterns of change in the α*E*Ψ lineage suggest significant functional constraint has operated for at least part of the time since α*E*Ψ arose. In this context we note some possible parallels in the evolution of the *Adh* pseudogene (*Adh*Ψ) in the *Drosophila repleta* radiation. Sullivan et al. (1994) used the distribution of shared disabling mutations across species to estimate the time of the inactivation of *Adh*Ψ. They then compared *Adh*Ψ sequences among species that had diverged after it had been inactivated and found that the rate of change was still constrained. Specifically they observed that there were fewer substitutions in the exons than in the introns, and both of these had fewer substitutions than the intergenic region. Furthermore, substitutions in *Adh*Ψ nonsynonymous sites are not significantly more frequent than nonsynonymous substitutions in an active *Adh* paralog. Either the timing of the inactivation event was incorrectly assigned (e.g., because the shared disabling mutations had independent origins) or there remained significant selective constraint on *Adh*Ψ after its inactivation. The latter explanation clearly conflicts with the traditional concept of pseudogene evolution. It will therefore be of interest to trace α*E*Ψ through the *Drosophila* radiation to determine when it was inactivated and to test for subsequent sequence constraint.

Acknowledgments. The authors wish to thank Richard Newcomb, Thomas Boyce, Charles Claudianos, Kerrie-Ann Smyth, Peter Campbell, and Dave Rowell for critical comments on the manuscript and valuable discussions. This work was funded in part by Australian Wool Research and Development Grants CEC12 and CEC71, Rural Industries Research and Development Grant CSE68A, and Horticultural Research and Development Grant HG420.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Beverley SM, Wilson AC (1984) Molecular evolution in *Drosophila* and the Higher Diptera. II A time scale for fly evolution. *J Mol Evol* 21:1–13

Cavener DR, Ray SC (1991) Eukaryotic start and stop translation sites. Nucleic Acids Res 19:3185–3192

Collet C, Nielsen M, Russell RJ, Karl M, Oakeshott JG, Richmond RC (1990) Molecular analysis of duplicated esterase genes in *Drosophila melanogaster.* Mol Biol Evol 7:9–28

Cygler M, Schrag JD, Sussman JL, Harel M, Silman I, Gentry MK, Doctor BP (1993) Relationship between sequence conservation and three dimensional structure in a large family of esterases, lipases, and related proteins. Protein Sci 2:366–382

Felsenstein J (1991) Phylip. University of Washington, Seattle

Field LM, Williamson MS, Moores GD, Devonshire AL (1993) Cloning and analysis of the esterase genes conferring insecticide resistance in the peach-potato aphid *Myzus persicae* (Sulzer). Biochem J 294:569–574

Genetics Computer Group (1994) GCG version 8.0. 575 Science Drive, Madison, WI 53711, USA

Gillespie JH, Kojima K (1968) The degree of polymorphism in enzymes involved in energy production compared to that in nonspecific enzymes in two *Drosophila ananassae* populations. Genetics 61:582–585

Graham GJ (1995) Tandem genes and clustered genes. J Theor Biol 175:71–87

Harel M, Sussman JL, Krejci E, Bon S, Chanal P, Massoulie J, Silman I (1992) Conversion of acetylcholinesterase to butyrylcholinesterase: modeling and mutagenesis. Proc Natl Acad Sci USA 89: 10827–10831

Healy MJ, Dumancic MM, Oakeshott JG (1991) Biochemical and physiological studies of soluble esterases from *Drosophila melanogaster.* Biochem Genet 29:365–388

Hughes PB, Raftos DA (1985) Genetics of an esterase associated with resistance to organophosphorus insecticides in the sheep blowfly, *Lucilia cuprina* (Wiedemann) (Diptera:Calliphoridae). Bull Entomol Res 75:535–544

Jentsch S, Seufert W, Sommer T, Reins H-A (1990) Ubiquitin-conjugating enzymes: novel regulators of eukaryotic cells. Trends Biochem Sci 15:195–198

Jones G, Venkataraman V, Ridley B, O'Mahony P, Turner H (1994) Structure, expression and gene sequence of a juvenile hormone esterase-related protein from metamorphosing larvae of *Trichoplusia ni.* Biochem J 302:827–835

Kumar S, Tamura K, Nei M (1993) MEGA: Molecular Evolutionary Genetics Analysis version 1.0. The Pennsylvania State University, University Park, PA 16802

Lotti M, Grandori R, Fusetti F, Longhi S, Brocca S, Tramontano A, Alberghina L (1993) Cloning and analysis of *Candida cylindracea* lipase sequences. Gene 124:45–55

Morton RA, Singh RS (1985) Biochemical properties, homology, and genetic variation of *Drosophila* ''non-specific'' esterases. Biochem Genet 23:959–972

Mouchès C, Pauplin Y, Agarwal M, Lemieux L, Herzog M, Abadon M, Beyssat-Arnaouty V, Hyrien O, de Saint Vincent BR, Georghiou GP, Pasteur N (1990) Characterization of amplification core and esterase B1 gene responsible for insecticide resistance in *Culex.* Proc Natl Acad Sci USA 87:2574–2578

Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C (1992) Splicing signals in *Drosophila:* intron size, information content, and consensus sequences. Nucleic Acids Res 20:4255–4262

Mutero A, Pralavorio M, Bride J-M, Fournier D (1994) Resistance associated point mutations in insecticide-insensitive acetylcholinesterase. Proc Natl Acad Sci USA 91:5922–5926

Oakeshott JG, van Papenrecht EA, Boyce TM, Healy MJ, Russell RJ (1993) Evolutionary genetics of *Drosophila* esterases. Genetica 90: 239–268

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A (1992) The α/β hydrolase fold. Protein Eng 5:197–211

Powell JR, DeSalle R (1995) *Drosophila* molecular phylogenies and their uses. In: Hecht MK (ed) Evolutionary biology. Plenum Press, p 88–137

Prestridge DS (1991) Signal scan; a computer program that scans DNA sequences for eukaryotic transcriptional elements. Comput Appl Biosci 7:203–206

Riley M (1989) Nucleotide sequence of the *Xdh* region in Drosophila pseudoobscura and an analysis of the evolution of synonymous codons. Mol Biol Evol 6:33–52

Ripoll DR, Faerman CH, Axelsen PH, Silman I, Sussman JL (1993) An electrostatic mechanism for substrate guidance down the aromatic gorge of acetylcholinesterase. Proc Natl Acad Sci USA 90:5128–5132

Rost B, Sander C (1993) Prediction of protein structure at better than 70% accuracy. J Mol Biol 232:584–599

Russell RJ, Robin GC, Kostakos P, Newcomb RD, Boyce TM, Medveczky KM, Oakeshott JG (1996) Molecular cloning of an a esterase gene cluster on chromosome 3R of *Drosophila melanogaster.* Insect Biochem Mol Biol 26:235–247

Sayle R (1994) RasMac molecular graphics. Modular CHEM Consortium

Schumacher M, Camp S, Maulet Y, Newton M, MacPhee-Quigley K, Taylor SS, Friedman T, Taylor P (1986) Primary structure of *Torpedo californica* acetylcholinesterase deduced from its cDNA sequence. Nature 319:407–409

Shrag JD, Cygler M (1993) 1.8A refined structure of the lipase from *Geotrichum candidum.* J Mol Biol 230:575–591

Smyth KA, Russell RJ, Oakeshott JG (1994) A cluster of at least three esterase genes in *Lucilia cuprina* includes malathion carboxylesterase and two other esterase genes implicated in resistance to organophosphates. Biochem Genet 32:437–453

Spackman ME, Oakeshott JG, Smyth K-A, Medveczky KM, Russell RJ (1994) A cluster of esterase genes on chromosome 3R of *Drosophila melanogaster* includes homologues of esterase genes conferring insecticide resistance in *Lucilia cuprina.* Biochem Genet 32:39–62

Sullivan DT, Starmer WT, Curtiss SW, Menotti-Raymond M, Yum J (1994) Unusual molecular evolution of an *Adh* pseudogene in *Drosophila.* Mol Biol Evol 11:443–458

Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I (1991) Atomic structure of acetylcholinesterase from *Torpedo californica:* a prototypic acetylcholine-binding protein. Science 253:872–879

Swofford DL (1993) PAUP: Phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign, IL

Swofford DL, Begle DP (1993) Paup manual. Lab. of Molecular systematics, Smithsonian Institute

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607

Vaughan A, Hemingway J (1995) Mosquito carboxylesterase Estα2 (A2). J Biol Chem 270:17044–17049

Vellom DC, Radic Z, Li Y, Pickering NA, Camp S, Taylor P (1993) Amino acid residues controlling acetylcholinesterase and butyrylcholinesterase specificity. Biochemistry 32:12–17

Villani F, White GB, Curtis CF, Miles SJ (1983) Inheritance and activity of some esterases associated with organophosphate resistance in mosquitoes of the complex of *Culex pipiens* L. (Diptera: Culicidae). Bull Entomol Res 23:154–170

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4691

Weller GL, Foster GG (1993) Genetic maps of the sheep blowfly *Lucilia cuprina:* linkage group correlations with other dipteran genera. Genome 36:495–506

Wirth MC, Marquine M, Georghiou GP, Pasteur N (1990) Esterases A2 and B2 in *Culex quinquefasciatus* (Diptera: Culicidae). Role in organophosphate resistance and linkage studies. J Med Entomol 7:202–206

Wright TRF, MacIntyre R (1963) A homologous gene/enzyme system. Esterase 6 in *Drosophila melanogaster* and *Drosophila simulans.* Genetics 48:1717–1726