A Cluster of Cuticle Protein Genes of *Drosophila melanogaster* at 65A: Sequence, Structure and Evolution

Jean-Philippe Charles,* Carol Chihara,† Shamim Nejad* and Lynn M. Riddiford*

**Department of Zoology, University of Washington, Seattle, Washington 98195-1800 and †Department of Biology, University of San Francisco, San Francisco, California 9411-1080*

Manuscript received March 27, 1997
Accepted for publication August 8, 1997

ABSTRACT

A 36-kb genomic DNA segment of the *Drosophila melanogaster* genome containing 12 clustered cuticle genes has been mapped and partially sequenced. The cluster maps at 65A 5-6 on the left arm of the third chromosome, in agreement with the previously determined location of a putative cluster encompassing the genes for the third instar larval cuticle proteins LCP5, LCP6 and LCP8. This cluster is the largest cuticle gene cluster discovered to date and shows a number of surprising features that explain in part the genetic complexity of the LCP5, LCP6 and LCP8 loci. The genes encoding LCP5 and LCP8 are multiple copy genes and the presence of extensive similarity in their coding regions gives the first evidence for gene conversion in cuticle genes. In addition, five genes in the cluster are intronless. Four of these five have arisen by retroposition. The other genes in the cluster have a single intron located at an unusual location for insect cuticle genes.

INSECT cuticle has a basic structure composed of a hydrophobic surface epicuticle and a fibrous inner procuticle made of an assembly of chitin and proteins. The large number of proteins comprising the cuticle and their importance in determining the physical characteristics of the cuticle are well known (ANDERSEN *et al.* 1995; WILLIS 1996 for reviews). Over the past few years, the sequences for a number of cuticle proteins and cuticle genes from various insects have been determined (ANDERSEN *et al.* 1995 for review), which should help our understanding of the molecular basis for the multiple and essential functions of this complex extracellular matrix.

The third larval instar cuticle of *Drosophila melanogaster* is a good model with which to approach this problem, as it contains only five major, and perhaps five minor proteins (FRISTROM *et al.* 1978) and has the advantage of ready genetic analysis. The genes encoding four of the major third larval instar cuticle proteins (LCP1–4) were isolated (SNYDER *et al.* 1981, 1982) and found to be clustered within 7.9 kb of genomic DNA at 44D on the second chromosome. Their genomic organization resembles that of the chorion genes (EICKBUSH and KAFATOS 1982; IATROU *et al.* 1982) where the genes are clustered in large arrays. A homologous cluster in *D. miranda* has a similar organization (STEINMANN and STEINMANN 1990). Since the characterization of the LCP1–4 cluster, two more cuticle gene clusters were found in the *D. melanogaster* genome, one at position 11 (CHIHARA and KIMBRELL 1986) and one at 84A on the third chromosome (FECHTEL *et al.* 1988; PULTZ 1988; PULTZ *et al.* 1988). Cuticle gene clusters have also been found in Lepidoptera (HORODYSKI and RIDDIFORD 1989) and Coleoptera (RONDOT *et al.* 1996), indicating that clustering of cuticle genes may be common in insect genomes.

In this article, we present the characterization of a cuticle gene cluster located at 65A on the left arm of the third chromosome of *D. melanogaster*. The mapping and sequencing data reveal a number of unexpected features. Twelve genes encoding a new family of Drosophila cuticle proteins are clustered within 22 kb, and the cluster is composed of two groups of divergently transcribed genes. This cluster contains multiple copy genes, and our data provide the first evidence that gene conversion has been a major driving force in the evolution of a cuticle gene family as was shown for the chorion protein genes (IATROU *et al.* 1984; EICKBUSH and BURKE 1985, 1986). The cluster includes also a new pseudogene and intronless genes that may have arisen by retroposition.

MATERIAL AND METHODS

Fly stocks: *D. melanogaster* were grown on standard agar-molasses-cornmeal-yeast media. The wild-type strain used for the cDNA library construction was Canton Special (Canton S). Sevelen, obtained from Dr. G. SCHUBIGER, University of Washington, is a wild-type strain that was originally collected in Zurich, Switzerland. The iso-1 strain (*y*[1]; *cn*[1] *bw* [1] *sp* [1]) is isogenic for all chromosomes (BRIZUELA *et al.* 1994) and was obtained from the Bloomington stock center. Oregon R wild type is described in CHIHARA and KIMBRELL (1986).

Nomenclature: Here we name the cuticle protein (cp) genes at 65A as *cp65A*, then designate the individual genes from left to right within the cluster by the letters a, b, c, etc.,

Corresponding author: Lynn M. Riddiford, Department of Zoology, University of Washington, Box 351800, Seattle, WA 98195-1800. E-mail: lmr@u.washington.edu

and the duplicated genes by number (see Figure 2). All but one of these genes are expressed in the larva (J.-P. CHARLES, C. CHIHARA, S. NEJAD and L. M. RIDDIFORD, unpublished data) and therefore are designated *LCP65Aa–g*; the one that is expressed only during adult development is designated *Acp65A*. For sake of simplicity in the text, we will refer to the genes as *Acp, Lcp-a, Lcp-b,* etc. In the case of *Lcp65A* genes corresponding to a previously described locus, the original locus name is indicated in parenthesis.

Isolation, mapping and sequencing of clones: A cDNA library was constructed in λgt10 from 5 μg polyA$^+$ RNA extracted from mixed early (0–6 hr) first and early (0–12 hr) whole second instar Canton S larvae using the Amersham cDNA library kit, and yielded 2.5×10^6 independent recombinants. The Oregon R cDNA library was prepared from early third instar larvae and was kindly provided by Dr. T. KAUFMAN, Indiana University. The iso-1 EMBL3 genomic library was obtained from Dr. J. TAMKUN (TAMKUN *et al.* 1992). Library screening, radioactive labeling of DNA probes, Southern hybridizations and preparation of DNA were performed using standard protocols (SAMBROOK *et al.* 1989), except as described below.

The EMBL3 clones were mapped by single or double digests followed by low or high stringency Southern hybridizations, as well as partial digests and indirect end-labeling as described (GOODE and FEINSTEIN 1990), except that the phage DNA was predigested with *Sma*I to remove the vector arms, and the digests were run in standard 0.7% agarose gels. The Canton S cDNAs were subcloned into the plasmid pBSK– (Stratagene) and were sequenced using the dideoxy-nucleotide chain termination method with [^{35}S]-dATP (Sequenase version 2.0 DNA sequencing kit; U.S. Biochemical). The Oregon R cDNA was amplified from a purified phage plug by PCR. Genomic subclones were sequenced mostly using the ABI Prism dye terminator cycle sequencing kit (Perkin Elmer).

Pairwise alignments were done with either the PILEUP program of Genetics Computer Group software package (DEVEREUX *et al.* 1984) or clustal w (THOMPSON *et al.* 1994). The phylogenetic analysis used the various programs of the PHYLIP package (Phylogeny Inference Package), version 3.5c (distributed by J. FELSENSTEIN, Department of Genetics, University of Washington, Seattle).

***In situ* hybridization to polytene chromosomes:** The cDNAs were biotinylated according to the procedure of HORODYSKI *et al.* (1989). Hybridization to Canton S salivary gland polytene chromosomes and detection were carried out according to the procedures for the Detek-I- HRP kit (ENZO diagnostics).

RESULTS

Genomic organization of the cuticle protein gene cluster at 65A: We prepared a cDNA library from RNA of first and second instar Drosophila of the Canton S strain and screened it under low stringency conditions with a 2.6-kb *Bgl*II-*Sal*I genomic fragment encompassing exons II–IV of the *Manduca sexta* LCP14 larval cuticle gene (REBERS and RIDDIFORD 1988). Among nine positive clones, one had a sequence identical to the sequence of the Drosophila LCP4 larval cuticle gene, which maps to the cuticle gene cluster at 44D (SNYDER *et al.* 1982). The other cDNAs encoded three proteins that were similar to the Manduca LCP14 protein and the Drosophila larval cuticle proteins (LCPs, SNYDER *et al.* 1982), and thus represented a new family of cuticle proteins. *In situ* hybridizations on polytene chromosomes using these cDNAs as probes and also a cDNA independently isolated from an Oregon R library showed that all of these genes map to the same region on the left arm of the third chromosome at 65A (data not shown). The position of the *Lcp-b* gene encoding LCP5 (see GenBank accession number U81550; J-P. CHARLES, C. CHIHARA, S. NEJAD and L. M. RIDDIFORD, unpublished data) was determined to be 65A 5-6 (Figure 1), thus coinciding with the predicted cuticle gene cluster at position 11 on this chromosome (CHIHARA and KIMBRELL 1986).

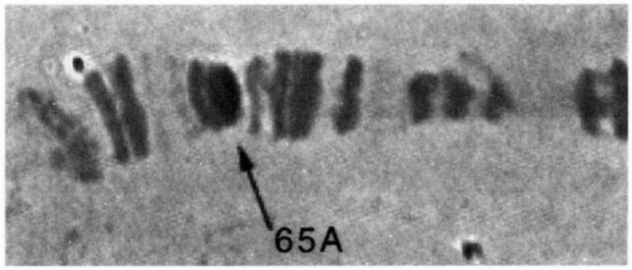

FIGURE 1.—*In situ* hybridization on Canton S polytene chromosomes with a *LCP65Ab* probe, showing a single band at 65A 5-6 on the left arm of the third chromosome.

Using our cDNAs as probes, we obtained 11 clones from a *D. melanogaster* iso-1 genomic library (TAMKUN *et al.* 1992). Four independent and overlapping clones were mapped and found to contain 12 cuticle protein genes or pseudogenes clustered within ~22 kb of genomic DNA (Figure 2A). The cluster is composed of two groups, seven (left group) and five (right group) tightly (average ~ 870 bp) spaced genes separated by a 4.5-kb spacer. The two groups are hereafter referred to as the left and right groups. All the genes of the left group, with the exception of *Acp,* are transcribed in the same direction (toward the left in Figure 2) and diverge from the right group genes.

Multiple copy genes: The left group contains a ~2.75-kb tandem repeat encompassing two genes (Figure 2A, duplication 1). The upstream regions of *Lcp-b1* and *-b2* are extensively conserved (Figure 3A). Importantly, a stretch of 595 bp spanning the entire open reading frame is identical in both genes. Consequently, the predicted protein products of the *Lcp-b1* and *-b2* genes should be identical. The location of the 5′ breakpoint of the duplication was not determined accurately, as the upstream sequences do not diverge significantly in this area. The presence of an *Eco*RV site upstream of each copy, however, suggests that the breakpoint lies near and upstream of these sites (see Figure 2B, duplication 1). The intergenic regions show a high level of identity through most of their length. Only nine base substitutions and insertions/deletions were observed over 750 bp, within the 1-kb sequence immediately downstream of the *Lcp-b* genes (~300 bp of the intergenic region were not determined).

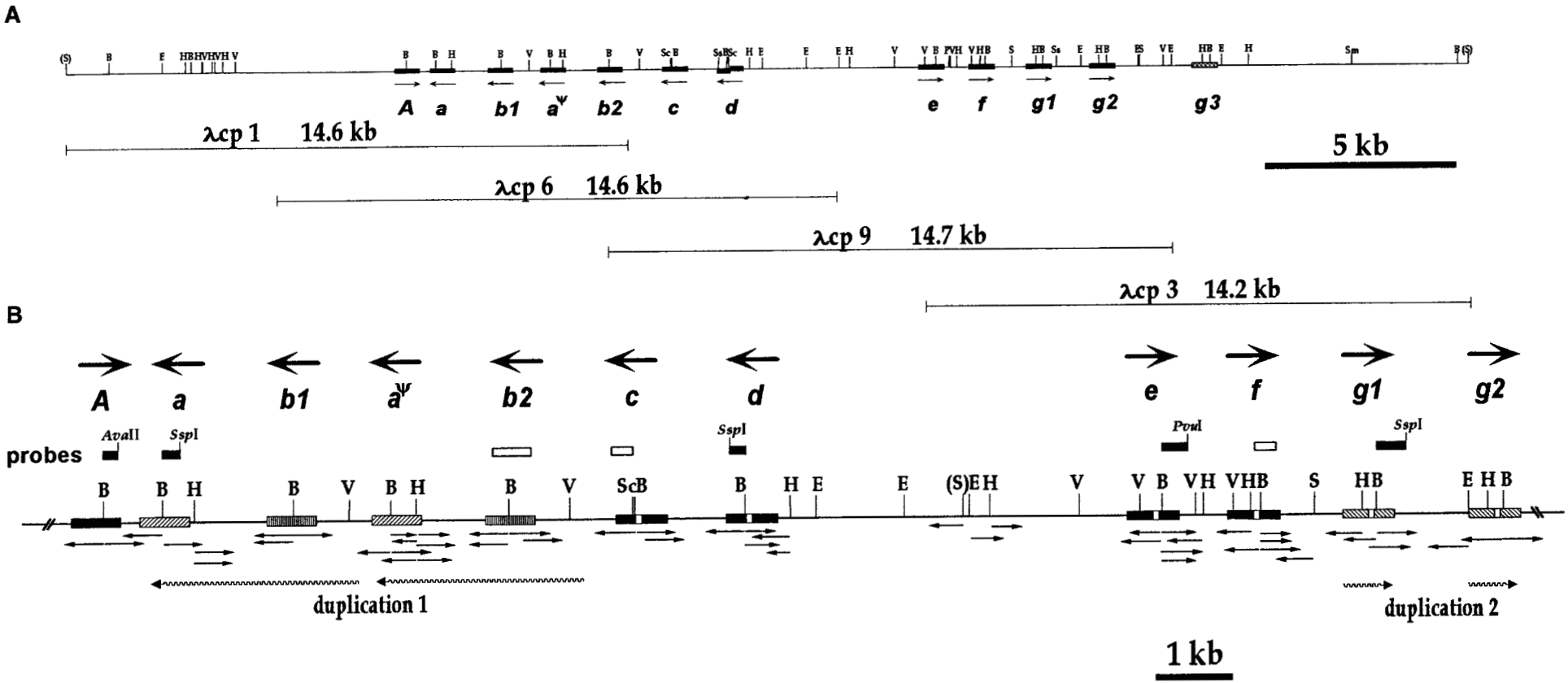

FIGURE 2.—Map of the 65A cuticle gene cluster. (A) The EMBL3 clones mapped are indicated below the restriction map of the area. Restriction sites are as follows: *Bam*HI (B), *Eco*RI (E), *Eco*RV (V), *Hin*dIII (H), *Sca*II (Sc), *Sal*I (S) and *Sma*I (Sm). *Sal*I sites corresponding to the EMBL3 cloning sites are indicated in parentheses. The map is complete for all sites but *Sca*II. All four lambda clones have been mapped for the six enzymes, with the exception of the ~3-kb rightmost *Sma*I-*Bam*HI fragment of λcp-3. The abbreviations for the genes are as follows: *A, Acp65A; a,LCP65Aa; b1, LCP65Ab1; b2, LCP65Ab2;* etc. (see text for details on nomenclature). The *LCP65Ag3* gene (*g3*) has been assigned to the position indicated in the map by Southern hybridization to the λcp-3 subclone, but has not been sequenced. The symbol Ψ is used to designate one of the two copies of the *LCP65Aa* gene (a^{Ψ}) as a pseudogene. This denotation is based on sequence data analysis presented in the text. The sequence data for these genes can be found in EMBL/GenBank under the following accession numbers: *Acp65A,* U8450; *Lcp65Aa,* U8448; *Lcp65Aa,* U8449; *Lcp65Ab1,* U8446; *Lcp65Ab2,* U844; *Lcp65Ac,* U8445; *Lcp65Ad,* U8324; *Lcp65Ae,* U8451; *Lcp65Af,* U8452; *Lcp65Ag1,* U8453; *Lcp65Ag2,* U8554. (B) Detail of the sequenced region. The direction of transcription is indicated on the top line; the probes used in this study [probes were from subcloned genomic fragments (■) or cDNA clones (□)] are on the second line. The arrows under the restriction map indicate the sequencing strategy. The boxes on the map correspond to transcription units, with introns indicated in white. Solid and striped boxes represent single and multiple copy genes, respectively. The wavy lines (duplications 1 and 2), respectively, show the location of a tandem repeat of the *LCP65Aa* and *LCP65Ab* genes and that of two copies (*LCP65Ag1* and *LCP65Ag2*) of the triplicated *LCP65Ag* gene.

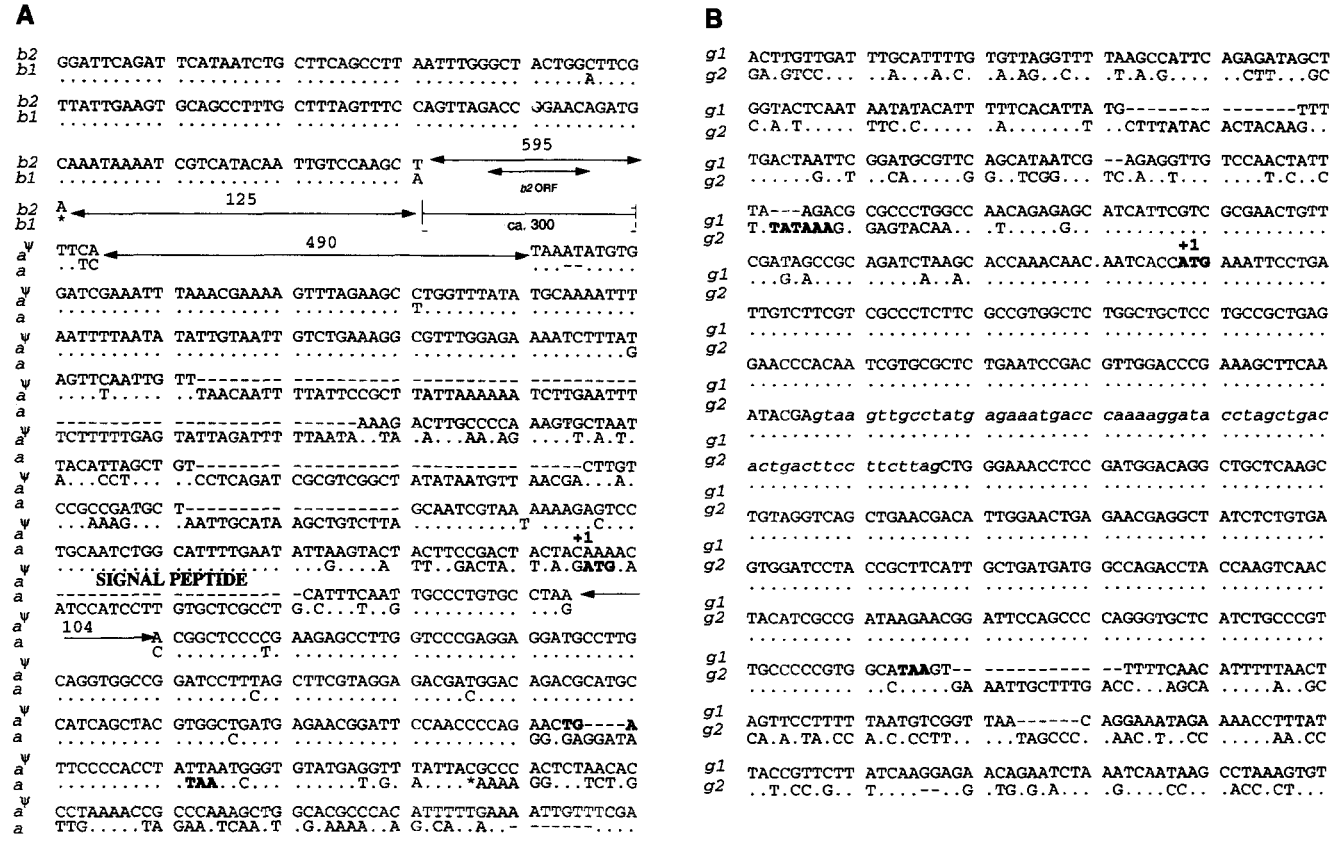

FIGURE 3.—Alignment of the duplicated regions. (A) Alignment of the two long tandem repeats of the left side of the cluster (duplication 1, in Figure 2B). The coding strands of the upstream repeat (*Lcp-b2* and *Lcp-a*$^{\Psi}$ genes, top line) and the downstream repeat (*Lcp-b1* and *Lcp-a* genes, bottom line) are aligned with the 5′ end on the top left in the figure. The gene sequences of each repeat are not contiguous, as an approximate 300-bp stretch of intergenic sequences (solid line with brackets) was not determined. Only those nucleotides differing from the *Lcp-b2/Lcp-a*$^{\Psi}$ repeat are reported on the bottom line. Nucleotides common to both sequences are indicated by dots. Solid lines with arrows indicate perfectly identical sequences, with the numbers indicating the exact length of the stretches. Dashes indicate deleted/inserted sequences. The 595-bp stretch encompasses the complete ORF for the *Lcp-b1* and *Lcp-b2* genes. Methionine initiator ATG and stop codons are in bold characters (+1 designates the first translated bp). (B) Alignment of the duplicated *Lcp-g1* and *Lcp-g2* genes. Conventions are the same as in A; intron sequences are in lowercase. There is a third (*Lcp-g3*) copy of gene *Lcp-g* in the cluster (see text) that was not sequenced and is not included in the figure.

In contrast, the sequences in the vicinity of the ATG initiation codon of the *Lcp-a* gene show a pattern of similarity in patches, with well conserved stretches interspersed with insertions/deletions of 19 to 64 bp. One of these is a 21-bp deletion in the reading frame of *Lcp-a*$^{\Psi}$ that removes part of the sequence coding for the signal peptide and shifts the reading frame. The *Lcp-a* and *Lcp-a*$^{\Psi}$ genes are highly similar over most of the remainder of the coding sequence. A 4-bp deletion in the hypothetical reading frame of *Lcp-a*$^{\Psi}$ introduces an early TGA stop codon, 19 bp before the TAA stop codon of the *Lcp-a* gene. The two sequences diverge completely 20 nucleotides thereafter, and no significant similarities were found over the next 300 bp of downstream sequences. The downstream breakpoint of the duplication therefore presumably lies a few base pairs downstream of the hypothetical stop codons of *Lcp-a*$^{\Psi}$.

The *Lcp-g* gene in the right group is likely triplicated. Two copies arranged in tandem (*Lcp-g1* and *-g2*) have been sequenced (Figure 3B), and the presence of a third copy on the right of *Lcp-g2* is indicated by high stringency-hybridization with λcp3 (*cf.* Figure 2A, data not shown) and genomic DNA (see below). The open reading frames in this duplication include a 61-bp intron and are identical, but for a single, conservative G/C transversion immediately before the stop codon. The sequences diverge greatly 60 bp upstream of the initiator ATG and 3 bp downstream of the TAA stop codon. These clear boundaries bracketing the open reading frame suggest that these two copies were homogenized through a gene conversion event (see DISCUSSION).

Multiple copy genes in other *D. melanogaster* strains: The presence of multiple copy genes in this cluster prompted us to look for repeated sequences in other *D. melanogaster* strains. The iso-1 strain was compared with two wild-type strains, Canton S and Sevelen, by genomic Southern analysis. Hybridization with the *Lcp-g1* probe showed three fragments of identical sizes in all three strains (Figure 4A). The sizes of these fragments match those predicted by the iso-1 map, thus indicating

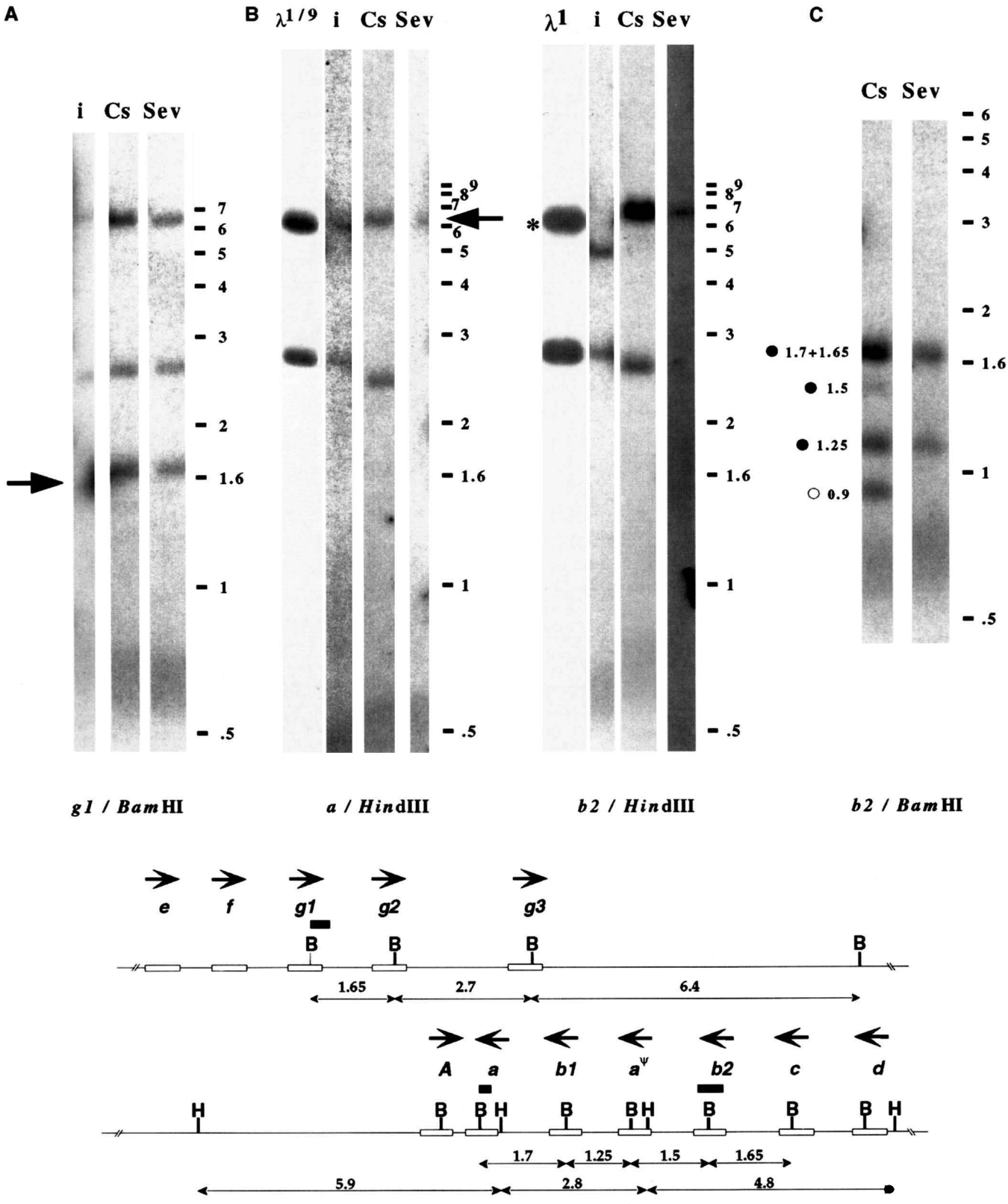

FIGURE 4.—Genomic southern blots of iso-1 (i), Canton S (Cs) and Sevelen (Sev) DNA. Each lane contains 2 μg of genomic DNA or 1 ng of lambda DNA (*cf.* Figure 2A) digested with restriction enzymes as indicated (λ^1, λcp-1; $\lambda^{1/9}$, 1:1 mixture of λcp-1 and λcp-9 DNA). The simplified maps at the bottom depict the positions of restriction sites and probes (■) pertinent to this study. The blots were hybridized with the *Lcp-g1* probe (A), *Lcp-a* (B) or *Lcp-b2* (B and C). The 1.65-kb iso-1 *Bam* HI (A) and the ~6.5-kb Sevelen *Hin*dIII (B) bands are faint, but clearly visible on the originals (arrows). The 4.8-kb *Hin*dIII genomic fragment of iso-1 encompassing *Lcp-b* (B right) is not represented in λcp1 (*cf.* Figure 2A), and the 6.2-kb fragment indicated (*) is generated by a *Hin*dIII cut in the right arm of EMBL3. In C the bands predicted by the iso-1 map are indicated by ● and the extra 0.9 kb detected in Canton S by ○.

```
Or.R                                    >G.................
isol  TCGTATAAAAGTCTCCACCCATCTGCACCAGAGCATCAAACAGTTCCAAG 50

cs-1        >CAGCTTT.T.TC.....ATG.................
cs-2.       >CAGCTTT.T.TC.....ATG.................
Or.R  ..........................ATG.................
isol  TTTTCTAACAAACACCACACAGCTCCAACATGAAATTCCTCATCGTCTTC 100

cs-1  ...................C.............................
cs-2  ...................C.............................
Or.R  ..........................................
isol  GTCGCCCTCTTCGCCATGGCAGTGGCCCGCCCCAACCTTGCCGAGATCGT 150

cs-1  ........................T.....................
cs-2  ........................T.....................
Or.R  ..........................................
isol  GAGGCAGGTCTCCGATGTTGAGCCCGAGAAGTGGAGCTCCGACGTGGAGA 200

cs-1  ..........................................
cs-2  ..........................................
Or.R  ..........................................
isol  CCAGCGATGGCACCAGCATCAAACAGGAGGGTGTCCTCAAGAACGCTGGC 250

cs-1  ..........................................
cs-2  ..........................................
Or.R  ..........................................
isol  ACTGACAACGAGGCCGCTGTCGTCCACGGATCCTTCACCTGGGTGGATGA 300

cs-1  ..........................................
cs-2  ..........................................
Or.R  ..........................................
isol  GAAGACCGGCGAGAAGTTCACCATCACATACGTGGCTGATGAGAACGGAT 350

cs-1  ....................................C..TAA..AT.A
cs-2  ....................................C..TAA..AT.A
Or.R  ....................................TAA......
isol  ACCAGCCCCAGGGCGCCCATCTGCCCGTGGCACCAGTTGCTTAAGATGTT 400

cs-1  .GATTG.CT..AATAAACT..T...T.GC.GC.AA.T.TAAAA..TGT.C
cs-2  .GATTG.CT..AATAAACT..T...T.GC.GC.AA.T.<A47
Or.R  ..............---.....................
isol  TTCCAAATCGATCAAA---GAGTTTAAAATAAATCAAAATGCTTTAAATT 450

cs-1  T....AATTCCTAATGT.GGATAAAA.G.AT...T...A..T.GCA.TGC
Or.R  <A8
isol  AAATTGGAGTAGCTATATTTCATGGTTCTTATATCTTCCCTCTATTTATA 500

cs-1  .AGAAAGACAATAAAT.T....CAA..C.<A28
isol  TTTGGTTGATTTGAACAAGATTTTTAATTGAAAAAAAAAAACCAGTTAGG 550
```

FIGURE 5.—Alignment of the iso-1 *Lcp-b1* gene (iso-1) with three independent cDNAs cloned from a Canton S library (Cs-1, Cs-2) and an Oregon R library (Or.R). Conventions are same as in Figure 3. The ATG initiator, stop codons and polyadenylation signals are in bold capital letters. Single nucleotide changes leading to amino-acid substitutions are also indicated in bold: the G-T (position 18) and G-C (position 389) transversions, respectively, lead to the substitution of Glu_{33} to Asp_{33} and Ala_{104} to Pro_{104} in the deduced Canton S polypeptide.

the presence of three copies of the *Lcp-g* gene organized in a similar way.

In contrast, the duplication in the left group shows restriction fragment length polymorphism. A single band was detected with either the *Lcp-a* or the *Lcp-b2* probe in *Hin*dIII-digested Sevelen DNA, whereas the two expected fragments were observed in iso-1 control DNA (Figure 4B). In addition, the *Lcp-b2* probe hybridized to only two fragments in *Bam*HI-digested Sevelen DNA (Figure 4C). Since this latter probe contains a conserved *Bam*HI site (and should therefore detect two bands for each *Lcp-b* copy), the data suggest that the *Lcp-a* and *Lcp-b* genes are not duplicated in the Sevelen strain.

In *Hin*dIII-digested Canton S DNA, both the *Lcp-a* and *Lcp-b* probes hybridized to two fragments differing from those of iso-1 DNA (Figure 4B). In *Bam*HI digests, all fragments corresponding to the 2.75-kb tandem repeat were present, but an additional 0.9-kb band was

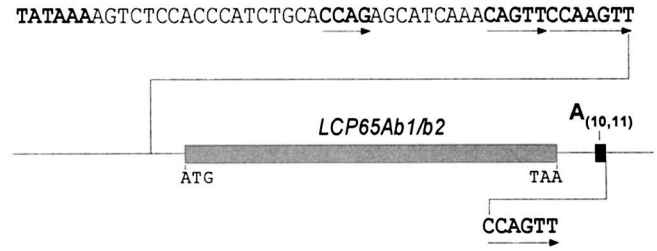

FIGURE 6.—Putative hallmarks of retroposition in the *Lcp65A-b* genes. ▨, open reading frame; ■, poly(A) tract. Three oligonucleotides (in bold) similar to the hexamer flanking the 3′ side of the poly(A) tract are found in direct orientation in the immediate 5′ flanking sequences of the *Lcp-b* genes (the first base of the TATA box is at −76 from the first coding base).

also detected (Figure 4C). When two different cDNA clones from our Canton S cDNA library (cs-1 and cs-2) were compared with the two iso-1 *Lcp-b* genes (Figure 5), the open reading frames were identical with the exception of three nucleotide substitutions, two of which lead to amino-acid substitutions (Glu_{33} to Asp_{33} and Ala_{104} to Pro_{104}). The untranslated regions, however, differed markedly, beginning 11 bp upstream of the ATG initiator and 3 bp downstream of the stop codon. By contrast, the sequence of an Oregon R cDNA was nearly identical with the iso-1 *Lcp-b* genes (Figure 5). These data indicate the possible presence of a third copy of the *Lcp-b* gene, *Lcp-b3*, in the Canton S strain (see DISCUSSION).

The structure of the 65A cuticle protein genes: The upstream and downstream regions (261 to 845 bp) of the cuticle protein genes (except for *Lcp-g3*) were sequenced and examined for the presence of consensus *cis* elements (not shown). All genes but *Lcp-g1* (lacking a TATA box) and *Lcp-a*$^{\psi}$ (with a frame-shifting deletion in the signal-peptide coding region and no consensus polyadenylation site) possess the *cis* elements expected from active genes.

The right group genes *Lcp-e, f, g* and the proximal left genes *Lcp-c* and *Lcp-d* have a single, small (58–91 bp) intron located ~60 bp downstream of the putative signal peptide cleavage site (Figure 2). In contrast, the *Lcp-b* cDNAs are colinear with the *Lcp-b* genes (*cf.* Figure 5), and the 5′ untranslated regions are very short and lack intron donor and acceptor consensus sites. The *Lcp-b* genes are thus most likely intronless. Inspection of the flanking regions of the *Lcp-b* genes revealed two additional features suggesting that the precursor gene might have arisen by retroposition. (1) A poly (A) stretch (of 11 and 10 bp, respectively) begins 133 bp downstream of the *Lcp-b1* and *Lcp-b2* stop codons. (2) Three short sequences (CCAG, CAGTT and CCAAGTT) that resemble the hexanucleotide CCAGTT flanking the poly(A) tract of these genes are found within 50 bp downstream of the TATA box (Figure 6).

FIGURE 7.—Alignment of the coding regions of the 65A cuticle genes. The coding strands of the 11 sequenced genes are shown with the ATG initiator codons aligned at the top left of the figure. Dots indicate positions with bases identical to those in the *Lcp-b2* gene. Intron sequences are in small letters, and gaps are shown in dashed lines. ATG and stop codons are underlined. Polymorphic sites specific to either the *Lcp-c*/*Lcp-d* gene pair, or the *Lcp-e, -f, -g* group are indicated by wavy underlines. The boxes show two shared polymorphisms between the *Lcp-c* gene and the *Lcp-e, -f, -g* group. See also text.

These sequences resemble the short direct repeats that typically flank retroposed sequences (WEINER 1986; BHANDARI *et al.* 1991).

The *Lcp-a* and *Lcp-a*$^{\Psi}$ genes also likely lack introns since they align with the *Lcp-b* intronless genes without introducing major gaps (Figure 7) and have no conserved intron acceptor and donor sites. The *Acp* gene also lacks conserved splicing sites, but when aligned with the other genes, it contains sequences that extend into the intronic region of the *Lcp-c, -d, -e, -f,* and *-g* genes (Figure 7).

Since the *Lcp-a*$^{\Psi}$ gene has a deletion in the putative signal peptide coding sequence and no in-frame ATG initiator codon, we conclude that this gene cannot give rise to a functional protein and is therefore most likely a pseudogene. All other genes possess a perfect open reading frame ranging from 297 bp (*Lcp-e*) to 327 bp (*Lcp-c*). Aside from the intronic and signal peptide regions described above, all sequences line up without gaps, with the exception of a 3-bp insertion/deletion (position 346 in Figure 7) and the 4-bp deletion near the putative stop codon of *Lcp-a*$^{\Psi}$ mentioned earlier. The most conserved region is a ~45-bp stretch that corresponds to the C-terminal domain of the proteins.

```
    1                              30                              60
d   MKFTIAIAFTCILACSLAAPPAIQQDAQVLRFDSD-VLPEGYKFAVETSDGKSHQEEGQL
c   ..C.V..V..ALF.VV....-.PDA.T.I..LE..-.Q....N..L......K.E.Q...
b1/2...L--.V.VALF.MAV.R.----NL.EIV..V..-.E..KWSSD......T.IKQ..V.
b3  ...L--.V.VALF.MAV.R.----NL.EIV..V..-.E.DKWSSD......T.IKQ..V.
a   ..SI--LVLA.L.SALCL.A--AAP..EIVDLV..-.NADS.SYKF.....TKQEQH.S.
A   --MLVVGSIALL..LAS.R.---QN.VE..EYESENTGLG....SYKL...T.RT...VV
e   ...L--.V.VAIF.FA..N-------E..IINLE..-.G..NFQWSF.....QAANAK...
f   ...L--.V.VALF.LAV.D-------V.I.K.E..-.G.VSFNYGY.....S.A.AA...
g   ...L--.V.VALF.VA....--AAEEPTIV.SE..-.G..SF.YDW.....QAA.AV...

   61                             90                             110
d   KDVGTDHEAIVVRGSYAYVGDD-GQTYSIQYLADENGFQPEGAHLPRPVQ
c   .NV..EQ.........SF.A..-....TVN.I.............NVPIGN
b1/2.NA...N..A..H..FTW.DEKT.EKFT.T.V.....Y..Q.....VAPVA
b3  .NA...N..A..H..FTW.DEKT.EKFT.T.V.....Y..Q.....VAPVP
a   .SL.PEED.LQ.A..FSF....-...HA.S.V........Q.EDI.HL
A   NNA...N.S.SI...VTW.AP.-....T.N.V.............K
e   .YPN....SLA.Q..FRF.A..-....EVN.I........Q.....VAS
f   .NV...E..LN.K.T.SF.A..-......A.T.....Y..Q.....VAPVV
g   N.I..EN...S.S...RFIA..-....QVN.I..K.....Q.....VAPVA
                G       D-G    V Y A   GY
                        E      L         F
                               I
```

FIGURE 8.—Alignment by the clustal w program of the Lcp65A and the Acp65A proteins deduced from the *cp65A* genes and cDNAs. Dots indicate positions of amino acids identical with Lcp-d. Dashes indicate gaps. The cuticle consensus sequence (ANDERSEN *et al.* 1995) is indicated underneath the sequences.

The open reading frames are then terminated after ~20 bp by either a TGA or TAA stop codon.

The cp65A genes encode a new family of hydrophilic cuticle proteins: The proteins deduced from the genes are aligned in Figure 8. These proteins are small (99–109 residues for the precursor proteins), hydrophilic polypeptides that are clearly homologous, and all have an hydrophobic leader-peptide, typical of secreted proteins. They line up with only a few gaps, with the interesting exception of the region corresponding to the cleavage site of the signal peptide that is quite variable. These gaps are due to small (3–12 bp) insertions or deletions in the signal-peptide coding regions (Figure 7). Also, there is a notably well conserved stretch of five residues ETSDG (positions 47–51 in Figure 8) in all but the Acp sequence that differs in the first two residues. This conserved sequence and variants are found in the *D. melanogaster* cuticle proteins LCP1, LCP2 (SNYDER *et al.* 1982), EDG78 and Gart (HENIKOFF *et al.* 1986; APPLE and FRISTROM 1991), and in larval cuticle proteins of the moths *Hyalophora cecropia* (hcp12: BINGER and WILLIS 1994) and *M. sexta* (LCP16/17: HORODYSKI and RIDDIFORD, 1989). Presumably these residues serve an important function in cuticular construction.

The C-terminal region is much more conserved, with a nearly perfectly conserved domain of 16 residues that includes the cuticle consensus (ANDERSEN *et al.* 1995). This domain, originally noticed by REBERS and RIDDIFORD (1988), has been found in a number of cuticle proteins from various arthropods and is generally hypothesized to be a binding site for chitin. The Lcp-b1, -b2, and -b3 proteins have an insertion in the middle of this domain and represent thus an exception to the consensus.

Table 1 shows the percentage identity and similarity values between the 65A genes derived from the alignments in Figures 7 and 8. The proteins have 32–100% identity (51% on average) and thus clearly form a new small cuticle protein family (for comparison, they show at most 26% identity with the proteins clustered at 44D). Genes *Lcp-a*$^{\psi}$ and *Acp* diverged most from the other genes in the cluster showing only an average 47% identity with the other genes. In addition to the obvious pairs of duplicated genes, the four genes in the right half of the cluster (*Lcp-e, -f, -g1,* and *-g2*) were found to form a distinct group ($P < 0.05$) by bootstrap analysis using parsimony maximum likelihood algorithms (not shown) (FELSENSTEIN 1985). The *Lcp-c/Lcp-d* pair (both genes and proteins) showed the highest levels of pairwise identity in the cluster and appeared in 85% of the trees in our analysis.

As expected from the fact that the *Lcp-c/Lcp-d* pair and the *Lcp-e, -f,* and *-g* genes form two phylogenetically distinct groups, a number of polymorphic positions involving several consecutive base pairs are found that are specific to one or the other group. A few examples of these are shown in Figure 7 (wavy lines). Because it is apparent that large sequence exchanges have occurred between the several duplicated genes in this cluster, we have looked in more detail for evidence of less conspicuous conversion events between genes. The number of polymorphic sites in Figure 7 is too large to allow a straightforward statistical analysis, but we note two examples of possible short conversion events between the *Lcp-c* gene and either *Lcp-e, -f,* or *-g* (boxes in Figure 7) that likely occurred at some time(s) after the segregation of the two groups. Alternatively, these changes could have been generated by other mechanisms, such as parallel mutation or double crossing over. Both however seem unlikely, and we think that these shared polymorphisms are best explained by the occurrence of conversion events.

DISCUSSION

The third instar cuticle of *D. melanogaster* contains six major (LCP1–6) and at least four minor (LCP7–10) urea-soluble proteins (FRISTROM *et al.* 1978; CHIHARA *et al.* 1982). LCP1–4 are encoded by four genes clustered within 7.9 kb of genomic DNA that maps to 44D (SNYDER *et al.* 1982). On the basis of meiotic mapping data (CHIHARA and KIMBRELL 1986), genes for LCP5, LCP6 and LCP8 were hypothesized to be clustered at position 11 on the third chromosome. The gene cluster at 65A that we describe here coincides well with their data. Further comparison of gene and protein sequences show that LCPs 5 and 6 are encoded by *LCP65Ab1/2* and LCP 8 by *LCP65Ag1/2/3* (J-P. CHARLES, C. CHIHARA, S. NEJAD and L. M. RIDDIFORD, unpublished data). Here we will analyze the structure and the evolution of this cluster of genes.

The multiple copy genes: Two of the genes within

TABLE 1

Table of identity/similarity between the 65A cuticle genes and the deduced proteins

	Lcp-b2	*Lcp-b1*	*Lcp-a*$^{\psi}$	*Lcp-a*	*Acp*	*Lcp-d*	*Lcp-c*	*Lcp-e*	*Lcp-f*	*Lcp-g1*	*Lcp-g2*
Lcp-b2		100/100[a]	NA	41/58	43/59	43/64	45/68	50/67	57/73	54/70	54/70
Lcp-b1	100[a]		NA	41/58	43/59	43/64	45/68	50/67	57/73	54/70	54/70
Lcp-a$^{\psi}$	50	50		NA[b]	NA	NA	NA	NA	NA	NA	NA
Lcp-a	56	56	83[b]		32/57	37/63	41/68	40/62	41/66	41/64	41/64
Acp	47	47	42	46		44/66	43/62	37/64	40/65	39/64	39/64
Lcp-d	55	55	45	49	52		65/80[c]	46/66	47/64	47/66	47/66
Lcp-c	61	61	50	55	52	71[c]		52/66	54/68	57/68	57/68
Lcp-e	61	61	49	54	46	55	60		63/80[d]	63/81[d]	63/81[d]
Lcp-f	65	65	48	52	46	56	61	69[d]		63/80[d]	63/80[d]
Lcp-g1	62	62	47	52	47	58	64	70[d]	70[d]		100/100[d]
Lcp-g2	62	62	47	52	47	58	64	70[d]	70[d]	99[d]	

The numbers are percentages. The bottom of the table shows the identity at the DNA level over the open reading frames (introns not included). The beginning of the reading frame of the *Lcp-a*$^{\psi}$ pseudogene was arbitrarily assigned to the position corresponding to the first codon of the *Lcp-a* gene (see Figure 7). The top part shows, respectively, the identity and similarity (estimated with the blosum30 matrix) values of the deduced proteins. Values with the same superscript indicate a monophyletic group. NA, not applicable.

the 65A cluster have duplicated or triplicated in the iso-1 line, and one of these, *Lcp-b,* appears to have one to three copies, depending on the strain (one in Sevelen, two in iso-1, and possibly three in Canton S). Multiple copy genes are quite commonly encountered among genes encoding proteins required in large amounts by the cell and are thought to arise from repeated DNA duplications (LI 1983). Moreover, copy number is also variable in clustered gene families such as the high cysteine chorion protein genes of *Bombyx mori* (YUE *et al.* 1988), usually as a result of unequal crossing over. In the 65A cluster, the two arrays of five and six tandemly arranged genes would clearly favor the occurrence of unequal crossing overs, thereby increasing the likelihood of gene duplications and copy number variation between strains.

The two Canton S cDNAs encoding the Lcp-b protein differ from the iso-1 genomic DNA and Oregon R cDNA by only three bases in the coding region, but diverge completely in both the 5′ and 3′ untranslated regions. These two cDNAs differ only by the length of their 3′ untranslated region and therefore likely correspond to the alternative use of the two polyadenylation signals. Although the differences between the iso-1 gene and the Canton S cDNA sequence could simply be allelic, we favor the hypothesis that the gene has duplicated again in the Canton S strain for the following reasons: (1) The Canton S cDNAs for the *Lcp-c* and *Lcp-f* genes were found to be identical (except for the intervening sequences) to the corresponding iso-1 genes (not shown; the GenBank accession numbers for these two cDNAs are U8445 and U8452, respectively). Thus, there is only a low level of polymorphism in these two strains. (2) The Oregon R Lcp-b cDNA is nearly identical to the iso-1 genomic DNA in both the coding and the noncoding regions, demonstrating that the noncoding sequences do not necessarily diverge rapidly. The confirmation of this hypothesis will require more detailed knowledge of the organization of the cluster and the sequences of the genes in other strains.

Sequence exchanges between the 65A genes: Members of multigene families often show a much greater degree of identity than would be expected if they were evolving independently. This process is known as concerted evolution and can be driven either by nonreciprocal recombination, *i.e.,* gene conversion (BALTIMORE 1981) or recurrent unequal sister chromatid exchanges (OHTA 1980). The *Lcp-a, Lcp-b,* and *Lcp-g* genes show an unusual degree of similarity with their respective copies (up to 595 consecutive identical bp between the *Lcp-b1* and *Lcp-b2* genes), a feature that can hardly be accounted for in terms of selection for structure. This high degree of positional identity is not unprecedented, as there have been similar findings in various kinds of organisms since the original study of the duplicated human $^{G}\gamma$ and $^{A}\gamma$ globin genes (SLIGHTOM *et al.* 1980). Examples include the rbcS-4 and rbcS-5 genes encoding the rubisco small subunit of *Mesembryanthemum crystallinum,* which are identical over 930 bp including the two introns (DEROCHER *et al.* 1993), and the winter flounder in which the 2A-b and 2A-c tandem repeat genes for antifreeze proteins share an identical stretch of 608 bp (DAVIES 1992). Similar extensively conserved stretches of DNA have also been found in two tandemly repeated collagen genes of *Caenorhabditis elegans* (PARK and KRAMER 1990) and in genes for proteins of the von Ebner's glands of rats (KOCK *et al.* 1994). The extremely high degree of identity between copies of the *Lcp65A* genes, coupled with the fact that some of the sequences involved (such as the *Lcp-a*$^{\Psi}$ pseudogene or the introns in the *Lcp-g1* and *-g2* genes) are not subject to stringent selection, is much more suggestive of gene conversion than of recurrent unequal sister chromatid exchanges.

The second type of evidence supporting this view

comes from the pattern of variation. For both the *Lcp-a* and *Lcp-g* genes, there is a sharp transition between identical (or nearly identical) coding regions and dissimilar flanking sequences. Similarly, the coding sequences of the iso-1 *Lcp-b1* and *-b2* genes and that of the Canton S *Lcp-b3* cDNAs are almost identical but diverge only 10 bp upstream of the initiator methionine codon and immediately after the stop codon. These features can hardly be explained by unequal crossovers, and similar patterns have been observed in genes involved in gene conversion. For instance, the coding sequences are identical or nearly so in the *M. crystallinum rbcS-4* and *rbcS-5* genes (DEROCHER *et al.* 1993) and the *C. elegans col-12* and *col-13* cuticle collagen genes (PARK and KRAMER 1990), whereas the upstream and downstream regions diverge greatly.

Conversion tracts have homology requirements (DENG and CAPECCHI 1992; SUGAWARA and HABER 1992; NASSIF and ENGELS 1993). Coding regions are under selective pressure and accumulate significantly fewer mutations than do flanking sequences. They are therefore more likely to support the elongation of conversion tracts. Such a combined action of selection and gene conversion, as noted earlier by PARK and KRAMER (1990) and HIBNER *et al.* (1991), can readily explain the restriction of homology to the coding sequences. The average length of meiotic conversion tracts in *D. melanogaster* is 352 bp (HILLIKER *et al.* 1994), which is the approximate length of the coding region of a cuticle gene. As noted by PARK and KRAMER (1990), the conservation of introns would depend primarily on the frequency of the conversion events. Similarly, shorter introns are less likely to hinder conversion tracts because they would accumulate fewer mismatches than large introns. The duplicated *Lcp-g* and *C. elegans* collagen genes (PARK and KRAMER 1990) have indeed perfectly conserved, very small (61–52 bp) introns. In contrast, the conversion tracts in the *Bombyx mori* chorion *ErA.1, ErA.2* and *ErA.3* genes are precisely limited on their 5′ end by much longer (375–880 bp) introns that show little sequence similarity (HIBNER *et al.* 1991).

Introns in cuticle genes: Genes *Lcp-c* and *-d* in the left region and *Lcp-e, -f,* and *-g* in the right region all have one intron and are more similar to each other than to *Lcp-a* and *-b* and *Acp*. Interestingly, the cuticle genes in the 44D cluster also have one intron but at a different position than those in the 65A cluster. The most parsimonious hypothesis is that the ancestral Drosophila cuticle gene had at least two introns, one of which was lost in the evolution of the 44D cluster and the other in the evolution of the 65A cluster. Furthermore, the position of the intron in the *Lcp65A* genes is the same as that of the second intron in the larval cuticle genes of the lepidopterans, *H. cecropia* (hccp12) (BINGER and WILLIS 1994) and *M. sexta* (mslcp 16/1, HORODYSKI and RIDDIFORD 1989; mscp14.6, REBERS *et al.* 1997). The position of the first intron in these lepidopteran genes is typical of that seen in the genes of the Drosophila 44D cluster. Although the sample size is small, these similarities suggest that the ancestral gene for cuticle genes of both Lepidoptera and Diptera had at least two introns.

The *Lcp-b* genes are intronless. One way to generate an intronless gene is through retroposition (ROGERS 1983). Retroposons are distinct from transposons and retroviruses and are abundant in both mammalian (ROGERS 1983, 1985) and insect (ADAMS *et al.* 1986) genomes. Both the *Lcp-b1* and *Lcp-b2* genes possess putative hallmarks of retroposition, *i.e.,* the lack of introns, the presence of a poly (A) tract at the 3′ end, and the presence of short flanking direct repeats (VANIN 1985; WEINER 1986). The conservation of these features following the duplication that led to the two *Lcp-b* genes suggests that the retroposition must have been relatively recent. To our knowledge there is no other known mechanism that might explain the lack of intron in a gene that is clearly homologous to intron-containing genes. Intron mobility is now well established for homing endonucleases (DOOLITTLE 1993), but no examples are known in Metazoa. That known mechanisms of DNA recombination such as conversion or crossing over could mimic the specificity of the splicing machinery also seems unlikely. We therefore conclude, in the absence of a better explanation, that both the *Lcp-a* and *Lcp-b* genes have most likely originated from the retroposition of a mature mRNA.

At the DNA and protein levels, *Lcp-b* is closest to the intron-containing gene *Lcp-f* and thus might have arisen from a processed *Lcp-f* mRNA. Most retroposons present genetic lesions or are not under the control of active promoters and thus do not contribute to the synthesis of active proteins (VANIN 1985). In mammals only a small number of functional retroposed genes are known (see references in BHANDARI *et al.* 1991; LONG and LANGLEY 1993; PERSSON *et al.* 1995). Previous to this study, the only functional retroposed insect gene known was the Drosophila *jingwei* gene (LONG and LANGLEY 1993). This gene arose by retroposition from the *alcohol dehydrogenase* gene followed by recruitment of additional 5′ exons and introns of an unrelated gene. The *Lcp-b1* and/or *Lcp-b2* genes are active in producing mRNAs and a protein(s) (J-P. CHARLES, C. CHIHARA, S. NEJAD and L. M. RIDDIFORD, unpublished data), so clearly are also functional retroposed genes.

The alignment of the *Acp* gene to the intron-containing genes suggests that although probably intronless, the *Acp* gene might have lost its intron by a process different from reverse transcription of a mature mRNA. The mechanisms involved could be an imperfect crossing over between two intron-containing genes, leading to the incorporation of part of an intron into the reading frame of the resulting chimera. One could also invoke an abortive conversion event between an intron-containing and an intronless gene. There is however

no phylogenetic evidence that the *Acp* gene is more closely related to the intronless genes than to the other genes. It is in fact more similar to *Lcp-c* and *-d* genes (see Table 1), but the low levels of identity observed clearly do not allow any conclusions on the origins of the *Acp* gene at present.

Origin and evolution of the 65A cluster: Our analysis of the data using different methods did not allow the construction of an unambiguous phylogeny for the 65A cluster. Two groups of sequences were found to be significant ($P < 0.05$) in most analyses and are presumably monophyletic. One group is comprised of the *Lcp-e, -f,* and *-g* genes on the right side of the cluster, the other of the *Lcp-c* and *-d* genes on the proximal left side. The cluster thus presumably arose from a gene of one of these two groups, and it is not clear from the data which was the ancestor. After one initial or a few duplications, a member of this ancestral group probably underwent duplication/inversion and "seeded" the other side of the cluster.

Another likely event in the history of the cluster is the reverse transcription of a fully processed mRNA and the insertion of the cDNA in the vicinity of the other genes. As such phenomena are not very common, the most parsimonious hypothesis is that it happened only once and that at least the two genes (*Lcp-a* and *Lcp-b*) that show a clear lack of intron sequences arose from a single retroposition event. The higher identity of the *Lcp-b* genes to the *Lcp-f* gene in the right group (see Table 1) suggests that the latter might be the founder gene. This hypothesis implies that the poly(A) tract resulting from the retroposition event of *Lcp-b* has not significantly mutated since the duplication that produced the *Lcp-a* and *Lcp-b* genes. The conservation of the poly(A) tract and other noncoding sequences for many generations would be expected to result, at least partly, from active selection on these sequences, which seems rather unlikely. We thus favor either one of the two following explanations: (1) The observed poly(A) and target site duplication might be simply fortuitous. In that case, presumably both genes arose from a single event. (2) Alternatively, the retroposition of *Lcp-b* might have been a very recent event, distinct from the retroposition of *Lcp-a*. The available data do not allow discrimination between these two possibilities, and obviously detailed genomic sequence information from other Drosophila strains is needed to resolve the issue.

Part of the difficulty of elucidating the phylogeny of the cluster presumably is owed to the shortness of the sequences compared. We suspect also that gene conversion events between genes other than the duplicated *Lcp-a, Lcp-b,* and *Lcp-g* genes have occurred at several occasions during the evolution of the cluster, thus obscuring the phylogenetic relationship among the different genes. Two examples of shared polymorphisms were found between the *Lcp-c* gene and the *Lcp-e, Lcp-f,* and *Lcp-g* genes that correspond likely to short conversion tracts. Interestingly, numerous short conversion events have played a major role in the evolution of a number of multigene families such as the *HcA* and *HcB* genes of *Bombyx mori* (EICKBUSH and BURKE 1985, 1986) and the α-amylase genes of *D. melanogaster* (INOMATA *et al.* 1995). Similarly, short and frequent gene conversions between a unique rearranged variable region and a pool of pseudogenes generates the immunoglobulin repertoire in birds (REYNAUD *et al.* 1978). Thus, gene conversion has been and still is a major driving force in the evolution of the Drosophila 65A cuticle gene cluster. This new family of genes could be an useful model to investigate the details of gene conversion in higher eucaryotes.

We thank WANDA MOATS for her help with the Drosophila cultures, Dr. KOSTAS IATROU for encouragement and advice, Dr. SCOTT EDWARDS for helpful discussions on the evolutionary aspects and a critical reading of the manuscript, and Dr. MICHAEL ASHBURNER for assistance in assigning the nomenclature for this cluster. This work was supported by National Science Foundation grants IBN 9100463 and IBN941995 to L.M.R. C.J.C. thanks TODD LAVERTY for teaching the chromosome spreading technique and is particularly indebted to the students who participated in much of the research: GREG SCHNEIDER, SHARON JIANG, PADMAJA MANDALAPARTHY, CHRISTIAN WADE and MARIO PINEDA. C.J.C. was supported in the early years of this work by a Bristol Myers Squibb Company Grant of Research Corporation and more recently by the Lily Drake Cancer Research Fund of University of San Francisco.

LITERATURE CITED

ADAMS, D. S., T. H. EICKBUSH, R. J. HERRERA and P. M. LIZARD, 1986 A highly reiterated family of transcribed oligo(A)-terminated, interspersed DNA elements in the genome of *Bombyx mori*. J. Mol. Biol. **18:** 465–478.

ANDERSEN, S. O., P. HØJRUP and P. ROEPSTORFF, 1995 Insect cuticular proteins. Insect Biochem. Mol. Biol. **25:** 153–176.

APPLE, R. T., and J. W. FRISTROM, 1991 20-hydroxyecdysone is required for, and negatively regulates, transcription of Drosophila pupal cuticle protein genes. Dev. Biol. **146:** 569–582.

BALTIMORE, D., 1981 Gene conversion: some implications for immunoglobulin genes. Cell **24:** 592–594.

BHANDARI, B., W. J. ROESLER, K. D. DELISIO, D. J. KLEMM, N. S. ROSS *et al.*, 1991 A functional promoter flanks an intronless glutamine synthetase gene. J. Biol. Chem. **266:** 7784–7792.

BINGER, L. C., and J. H. WILLIS, 1994 Identification of the cDNA, gene and promoter for a major protein from flexible regions of the giant silkmoth *Hyalophora cecropia*. Insect Biochem. Mol. Biol. **24:** 989–1000.

BRIZUELA, B. J., L. ELFRING, J. BALLARD, J. W. TAMKUN and J. A. KENNISON, 1994 Genetic analysis of the *brahma* gene of *Drosophila melanogaster* and polytene chromosome subdivisions 2AB. Genetics **137:** 803–813.

CHIHARA, C., and D. A. KIMBRELL, 1986 The cuticle proteins of *Drosophila melanogaster*: genetic localization of a second cluster of third-instar genes. Genetics **114:** 393–404.

CHIHARA, C., D. J. SILVERT and J. W. FRISTROM, 1982 The cuticle proteins of *Drosophila melanogaster*: stage specificity. Dev. Biol. **89:** 379–388.

DAVIES, P. L., 1992 Conservation of antifreeze protein-encoding genes in tandem repeats. Gene **112:** 163–170.

DENG, C., and M. R. CAPECCHI, 1992 Reexamination of gene targeting efficiency as a function of the extent of homology between the targeting vector and the target locus. Mol. Cell. Biol. **12:** 3365–3371.

DEROCHER, E. J., F. QUIGLEY, R. MACHE and H. J. BOHNERT, 1993 The six genes of the Rubisco small subunit multigene family from *Mesembryanthemum crystallinum*, a facultative CAM plant. Mol. Gen. Genet. **239:** 450–462.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

DOOLITLE, R. F., 1993 The comings and goings of homing endonucleases and mobile introns. Proc. Natl. Acad. Sci. USA **90:** 5379–5381.

EICKBUSH, T. H., and W. D. BURKE, 1985 Silkmoth chorion gene families contain patchwork patterns of sequence homology. Proc. Natl. Acad. Sci. USA **82:** 2814–2818.

EICKBUSH, T. H., and W. D. BURKE, 1986 The silkmoth late chorion locus. II. Gradients of genes conversion in two paired multigene families. J. Mol. Biol. **190:** 357–366.

EICKBUSH, T. H., and F. C. KAFATOS, 1982 A walk in the chorion locus of *Bombyx mori.* Cell **29:** 633–643.

FECHTEL, K., J. E. NATZLE, E. E. BROWN and J. W. FRISTROM, 1988 Prepupal differentiation of Drosophila imaginal discs: identification of four genes whose transcripts accumulate in response to a pulse of 20-hydroxyecdysone. Genetics **120:** 465–474.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 83–91.

FRISTROM, J. W., R. J. HILL and F. WATT, 1978 The procuticle of Drosophila: Heterogeneity of urea-soluble proteins. Biochemistry **1:** 3917–3924.

GOODE, B. L., and S. C. FEINSTEIN, 1990 Restriction mapping of recombinant λ DNA molecules using pulse field gel electrophoresis. Anal. Biochem. **191:** 70–74.

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

HIBNER, B. L., W. D. BURKE and T. H. EICKBUSH, 1991 Sequence identity in an early chorion multigene family is the result of localized gene conversion. Genetics **128:** 595–606.

HILLIKER, J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.,* 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster.* Genetics **13:** 1019–1026.

HORODYSKI, F. M., and L. M. RIDDIFORD, 1989 Expression and hormonal control of a new larval cuticular multigene family at the onset of metamorphosis of the tobacco hornworm. Dev. Biol. **132:** 292–303.

HORODYSKI, F. M., L. M. RIDDIFORD and J. W. TRUMAN, 1989 Isolation and expression of the eclosion hormone gene from the tobacco hornworm, *Manduca sexta.* Proc. Natl. Acad. Sci. USA **86:** 8123–8127.

IATROU, K., S. G. TSITILOU and F. C. KAFATOS, 1982 Developmental classes and homologous families of chorion genes in *Bombyx mori.* J. Mol. Biol. **15:** 417–434.

IATROU, K., S. G. TSITILOU and F.C. KAFATOS, 1984 DNA sequence transfer between two high-cysteine chorion gene families in the silkmoth *Bombyx mori.* Proc. Natl. Acad. Sci. USA **81:** 4452–4456.

INOMATA, N., H. SHIBATA, E. OKUYAMA and T. YAMAZAKI, 1995 Evolutionary relationships and sequence variation of α-amylase variants encoded by duplicated genes in the *Amy* locus of Drosophila melanogaster. Genetics **141:** 237–244.

KOCK, K., C. AHLERS and H. SCHMALE, 1994 Structural organization of the genes for rat von Ebner's gland proteins 1 and 2 reveals their close relationship to lipocalins. Eur. J. Biochem. **221:** 905–916.

LI, W.-H., 1983 Evolution of duplicate genes and pseudogenes, pp. 14–37 in *Evolution of Genes and Proteins,* edited by M. NEI and R. K. KOEHN. Sinauer, Sunderland, MA.

LONG, M., and C. H. LANGLEY, 1993 Natural selection and the origin of *jingwei,* a chimeric processed functional gene in *Drosophila.* Science **260:** 91–95.

NASSIF, N., and W. ENGELS, 1993 DNA homology requirements for mitotic gap repair in Drosophila. Proc. Natl. Acad. Sci. USA **90:** 1262–1266.

OHTA, T., 1980 *Evolution and Variation in Multigene Families: Lecture Notes in Biomathematics,* Vol. 3, Springer Verlag, New York.

PARK, Y.-S., and J. M. KRAMER, 1990 Tandemly duplicated *Caenorhabditis elegans* collagen genes differ in their mode of splicing. J. Mol. Biol. **211:** 395–406.

PERSSON, K., I. HOLM and O. HEBY, 1995 Cloning and sequencing of an intronless mouse S-adenosylmethionine decarboxylase gene coding for a functional enzyme strongly expressed in the liver. J. Biol. Chem. **20:** 5642–5648.

PULTZ, M. A., 1988 A molecular and genetic analysis of *Proboscipedia* and flanking genes in the *Antennapedia Complex* of *Drosophila melanogaster.* Ph.D. dissertation, Indiana University, Bloomington, IN.

PULTZ, M. A., R. J. DIEDERICH, D. L. CRIBBS and T. C. KAUFMAN, 1988 The *proboscipedia* locus of the *Antennapedia-Complex*: a molecular genetic analysis. Genes Dev. **2:** 901–920.

REBERS, J. E., and L. M. RIDDIFORD, 1988 Structure and expression of a *Manduca sexta* larval cuticle gene homologous to Drosophila cuticle genes. J. Mol. Biol. **203:** 411–423.

REBERS, J. E., J. NIU and L. M. RIDDIFORD, 1997 Structure and spatial expression of the *Manduca sexta* MSCP14.6 cuticle gene. Insect Biochem. Mol. Biol. **2:** 229–240.

REYNAUD, C-A., V. ANQUEZ, H. GRIMAL and J. C. WEILL, 1987 A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell **48:** 379–388.

ROGERS, J. H., 1983 Retroposons defined. Nature **301:** 460.

ROGERS, J. H., 1985 The origin and evolution of retroposons. Int. Rev. Cytol. **93:** 18–29.

RONDOT, I., J. P. DELBECQUE and J. DELACHAMBRE, 1996 Structure and hormonal regulation of clustered cuticular protein genes in *Tenebrio molitor.* Abstracts, XII Ecdysone Workshop, Barcelona.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1987 *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

SLIGHTOM, J. L., A. E. BLECHL and O. SMITHIES, 1980 Human fetal $^{G}\gamma$ and $^{A}\gamma$ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21:** 627–638.

SNYDER, M., J. HIRSH and N. DAVIDSON, 1981 The cuticle genes of Drosophila: A developmentally regulated gene cluster. Cell **25:** 165–177.

SNYDER, M., M. HUNKAPILLER, D. YUEN, D. SILVERT, J. FRISTROM *et al.,* 1982 Cuticle protein genes of Drosophila: structure, organization and evolution of four clustered genes. Cell **29:** 1027–1040.

STEINMANN, M., and S. STEINMANN, 1990 Evolutionary changes in the organization of the major LCP gene cluster during sex chromosomal differentiation in the sibling species *Drosophila persimilis, D. pseudoobscura,* and *D. miranda.* Chromosoma **99:** 424–431.

SUGAWARA, N., and J. E. HABER, 1992 Characterization of double-strand break-induced recombination: homology requirements and single-stranded DNA formation. Mol. Cell. Biol. **12:** 563–575.

TAMKUN, J. W., R. DEURING, M. P. SCOTT, M. KISSINGER, A. M. PATTATUCCI *et al.,* 1992 Brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. Cell **68:** 561–572.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

VANIN, E. F., 1985 Processed pseudogenes: characteristics and evolution. Int. Rev. Genet. **19:** 253–272.

WEINER, A. M., 1986 Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55:** 631–661.

WILLIS, J. H., 1996 Metamorphosis of the cuticle, its proteins and their genes, pp. 253–282 in *Metamorphosis: Postembryonic Reprogramming of Gene Expression in Amphibian and Insect Cells,* edited by L. I. GILBERT, J. R. TATA and B. G. ATKINSON, Academic Press, San Diego.

YUE, X. N., B. SAKAGUCHI and T. H. EICKBUSH, 1988 Gene conversions can generate sequence variants in the late chorion multigene families of *Bombyx mori.* Genetics **120:** 221–231.

Communicating editor: S. HENIKOFF