# ORGANIZATION AND EXPRESSION OF A CLUSTER OF DROSOPHILA CUTICLE GENES

Thesis by

Michael Snyder

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1983

(Submitted August 31, 1982)

## ACKNOWLEDGEMENTS

I thank Norman Davidson whose outstanding scientific abilities and seemingly unlimited patience provided an excellent research environment. I also owe thanks to the members of Norman Davidson's group for, among other things, helping with a critical part of the DNA sequencing. My special appreciation to N. Davis Hershey for stimulating discussions, writing lessons, and creative encouragement. My thanks to the Biology Faculty for useful discussions. I am grateful to the students for capricious but sometimes relevant discourses, and also for valuable lessons on punctuality at afternoon sessions in the Delbrück lounge.

My final thanks to the Caltech secretaries for their excellent professional service time and time and time and time again.

## ABSTRACT

A 50 kb DNA segment of the Drosophila genome has been cloned and characterized. This segment lies at chromosomal location 44D and contains two small gene families. One family is comprised of four related cuticle genes clustered within 7.9 kb of DNA. The four genes encode four of the five major third instar larval cuticle proteins. These cuticle genes are coordinately expressed in the integument of third instar larvae, and they are not abundantly expressed in other developmental stages. A fifth cuticle-like gene lies within this gene cluster. It is judged to be a pseudogene, because several features of its structure and the absence of transcripts suggest that it is nonfunctional. Sequence comparisons indicate it arose by an unequal crossing over event involving two closely related and adjacent cuticle genes.

Eleven kb away from the cuticle gene cluster lies another gene family. This family is comprised of three genes that are 55-60% homologous in DNA sequence and clustered within 8 kb of DNA. The three genes are expressed together in larval stages and adults but show a different pattern of developmental expression from the third instar larval cuticle protein genes. Thus two small gene families can lie adjacent on the chromosome and exhibit different patterns of developmental expression, even though individual genes within a clustered family are coordinately expressed.

Additionally, a Drosophila strain has been studied which fails to synthesize one of the cuticle proteins. A molecular characterization of this strain is reported, which includes the finding of a transposable element in the promoter region of the unexpressed gene.

## TABLE OF CONTENTS

# CHAPTER 1

Cell, Vol. 25, 165-177, July 1981, Copyright © 1981 by MIT

# The Cuticle Genes of Drosophila: A Developmentally Regulated Gene Cluster

**Michael Snyder, Jay Hirsh*† and Norman Davidson***
Division of Biology and Department of Chemistry*
California Institute of Technology
Pasadena, California 91125

## Summary

**A 36 kilobase (kb) DNA segment of the Drosophila genome that contains several larval cuticle protein genes has been cloned and characterized. This segment maps at chromosomal locus 44D. It contains five genes, all of which are expressed at the same time of Drosophila development. Four of the genes are clustered within 7.9 kb of DNA and are abundantly expressed as poly(A) RNA in the epidermis of late third instar larvae but are not abundantly expressed in other developmental stages. A fifth gene lies 8 kb away from this cluster and is expressed at a much lower level in late third instar larval poly(A) RNA. Three of the four abundantly expressed genes have been shown to code for larval cuticle proteins; less decisive evidence indicates that the fourth gene also probably codes for a larval cuticle protein. Some of the genes are related in DNA sequence, and the proteins encoded in the cluster are related immunologically. Thus the cuticle genes encoded by the segment at 44D are members of a family of genes of common ancestry, which share the same pattern of developmental expression and reside in a small segment of the Drosophila genome.**

## Introduction

The mechanism by which eucaryotes coordinately express specific sets of genes is unknown. To understand the underlying processes, several systems in which a number of genes are coordinately expressed both in the same tissue and at the same time of development are now under investigation.

The larval cuticle genes of Drosophila represent such a system in that a small battery of structural genes is expressed in the epidermal cells of late third instar larvae. Five major cuticle proteins are synthesized and secreted by these cells (Fristrom et al., 1978). These proteins and chitin are major components of the cuticle. During formation of the pupa, cuticular components become covalently crosslinked to form a hard, brown, water-impermeable case, which surrounds and protects the animal during its subsequent development (for a review on cuticle synthesis, see Hepburn, 1976).

The larval cuticle genes are probably also an example of a battery of genes whose expression is hormonally induced. Evidence from many insects indicates that the molting hormone, ecdysone, induces larval cuticle synthesis. The well characterized enzyme, dopa decarboxylase, which is involved in cuticle synthesis exhibits an ecdysone-dependent pattern of expression in Drosophila and other insects (Fragoulis and Sekeris, 1975; Karlson and Sekeris, 1962; Chen and Hodgetts, 1974; Kraminsky et al., 1980). With the isolation of probes specific for the cuticle protein genes, it will be possible to perform direct tests of the effects of ecdysone on the expression of these genes.

We describe the cloning and characterization of several larval cuticle protein genes. One of our interesting results is that several of these coordinately expressed genes are clustered in a small segment of the Drosophila genome.

## Results

### The Larval Cuticle Proteins

When proteins are extracted from purified cuticles of late third instar larvae and studied by gel electrophoresis under nondenaturing conditions, five major species, denoted CP1 through CP5, are observed (Fristrom et al., 1978). We have further studied these proteins by both one-dimensional SDS gel electrophoresis (not shown) and standard two-dimensional gel electrophoresis (Figure 1A). As shown in Figure 1A, the five cuticle proteins are resolved on two-dimensional O'Farrell gels according to their molecular weights, which range from 9 kilodaltons for CP3 to 17.5 kilodations for CP1 and CP2. They are also separated by their isoelectric points, which are 5.9 for CP1 and 5.7 for CP2; CP3, CP4 and CP5 have isoelectric points in the range of 4.4-5.0. In addition to the major larval cuticle proteins, five minor species are also resolved. These are difficult to see in Figure 1, and we have not attempted to characterize them further. None of the major protein bands appears to contain carbohydrate residues, as indicated by the absence of staining with basic fuchsin stain following appropriate treatments (Keyser, 1964) (data not shown). The sensitivity of this assay was such that one or two glucose residues per protein molecule would have been detected.

### Selecting a Cuticle Clone

A recombinant phage with an insert coding for cuticle proteins was isolated by a four-step procedure. A random shear Drosophila recombinant DNA library (Maniatis et al., 1978) was screened with a cDNA probe made to poly(A) RNA isolated from the integument of late third instar larvae. The integument is enriched for epidermal cells that are actively synthesizing cuticle proteins (J. Hirsh and N. Davidson,

† Present address: Department of Biological Chemistry, Harvard Medical School, Boston, Massachusetts 02115

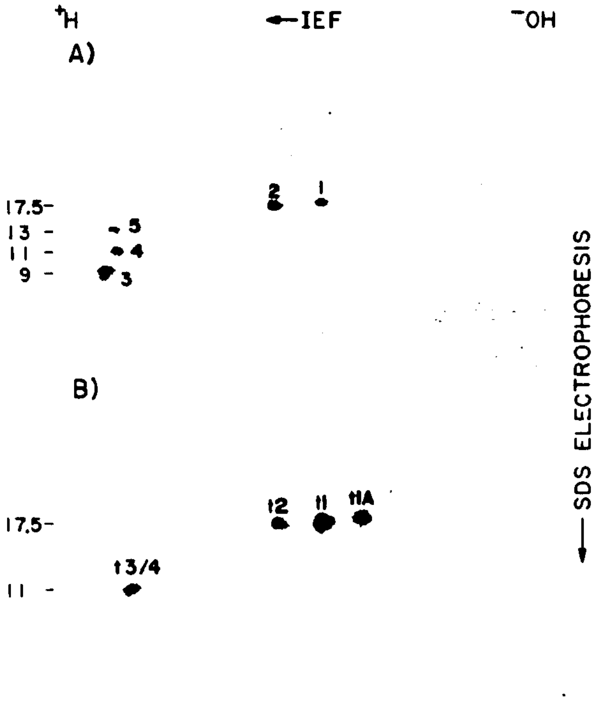


Figure 1. The Larval Cuticle Proteins and Proteins Translated From RNA Selected by λDmLCP1

(A) Larval cuticle proteins were isolated and analyzed on two-dimensional gels as described in Experimental Procedures. The gel was stained with Coomassie blue. (B) Autoradiogram of a two-dimensional gel containing labeled proteins translated from RNA selected by λDmLCP1. These proteins were mixed with unlabeled larval cuticle proteins as described in Experimental Procedures. Comparison of the Coomassie stains of this gel (not shown) with the autoradiogram spots shows that t1 overlaps exactly with CP1 and t2 with CP2; t3/4 overlaps with CP4 but is slightly more basic. Molecular weights in kilodaltons are indicated on the left side of the gels.

submitted). From 40,000 phages screened, 300 positive phages were selected. These were subsequently put through two cycles of counterselection.

Clones were screened for the relative intensity with which phage plaques hybridized to cDNA probes made to larval integument poly(A) RNA and to embryo poly(A) RNA, which is not expected to contain cuticle messages. We obtained 37 phages that hybridized more intensely to the larval integument probe. Of these, 16 were subjected to a second cycle of counterselection, in which the phage DNAs were tested for their differential hybridization to embryo, whole larval, pupal and larval integument cDNA probes. The four phage DNAs that gave the greatest hybridization signal to the larval integument probe relative to the other three were put through the final screening.

In this final selection, phage DNAs were used to positively select RNA from total late larval poly(A) RNA. This RNA was translated in vitro, and protein products were analyzed on two-dimensional O'Farrell gels. One clone, denoted λDmLCP1, selected RNA that upon translation yielded polypeptides with two-dimensional gel migration patterns similar to those of known cuticle proteins (Figure 1B; compare with Figure 1A). Polypeptides synthesized from the selected RNA were resolved into four spots, denoted t1A, t1, t2 and t3/4. This nomenclature is based on the identification of translation products with cuticle proteins as shown in the next section. Spots t1A, t1 and t2 have molecular weights of approximately 17.5 kilodaltons but differ in their isoelectric points. The polypeptide or polypeptides in the t3/4 spot have a molecular weight of 11 kilodaltons. Occasionally, a fifth protein spot is seen migrating next to t3/4. This spot may be due to streaking of t3/4. The finding that one phage selects RNAs that, when translated, give a number of proteins raised two questions: which of these polypeptides are, or are precursors to, larval cuticle proteins; and how many genes are encoded by the insert of this clone.

During this work, it became evident that an additional gene or genes lay on sequences flanking one side of λDmLCP1. To select overlapping clones, we constructed probes to the ends of the λDmLCP1 insert and screened a Drosophila recombinant DNA library, which contains partial Eco RI digests of Drosophila DNA (Yen et al., 1979; Davidson et al., 1980). From the series of overlapping clones obtained, one phage, denoted λDmLCP3, which overlapped 3 kilobases (kb) with λDmLCP1 and contained an additional 16 kb of flanking DNA including the sequences of interest, has been characterized. When the RNA coding segment of λDmLCP3 is used for the positive RNA selection and translation procedure (see below), no additional polypeptides other than those observed in Figure 1B are seen on two-dimensional gels.

## The In Vitro Translation Products Are Cuticle Proteins

Several assays were used to determine which of the in vitro translation products from RNAs selected by λDmLCP1 and λDmLCP3 are cuticle proteins or their precursors. The first was to test for which of the labeled in vitro translation products comigrate with unlabeled in vivo cuticle proteins in two-dimensional gels. Figure 1B shows the fluorogram from such an experiment. The translation product t1 comigrates with the known cuticle protein CP1 and t2 with CP2, whereas t1A has the same molecular weight but is more acidic. This evidence suggests that t1 is CP1 and t2 is CP2, and leaves t1A unassigned. The translation product t3/4 migrates to a position very close to cuticle proteins CP3 and CP4 (Figure 1B). Thus the t3/4 spot could contain a precursor or the mature polypeptide for either CP3 or CP4, or it could be a composite spot containing both polypeptides. We argue below that the latter hypothesis is more consistent with the data and show that none of these translation products is a precursor to CP5.

Immunoprecipitation tests also demonstrate that these polypeptides are cuticle proteins. Two types of antisera against larval cuticle proteins were supplied by D. Silvert and J. Fristrom. One, termed anti-LCP, was prepared against all five cuticle proteins and binds all five cuticle proteins. The other, termed anti-3, is from rabbits immunized with purified CP3 alone. However, anti-3 immunoprecipitates CP1, CP2, CP3 and CP4. Thus these four proteins share common antigenic determinants (D. Silvert and J. Fristrom, in preparation). As shown in Figure 2, part I, when polypeptides were translated from either total poly(A) RNA or from λDmLCP1-selected RNA and then immunoprecipitated with anti-3, the same bands of 17.5 and 11 kilodaltons on SDS gels were observed. No such protein bands can be found when preimmune serum (lanes b and d) is used or in translations where no exogenous RNA is added (lanes f and g). These same results were obtained with anti-LCP serum (not shown). To distinguish which 17.5 kilodalton proteins are recognized by the anti-3 serum, the translation nroducts of λDmLCP1-selected RNA and of total poly(A) RNA were immunoprecipitated and analyzed on two-dimensional gels. As shown in Figure 2B, all three 17.5 kilodalton polypeptides, t1A, t1 and t2, as well as the 11 kilodalton polypeptide, t3/4, bind to the anti-3 serum. Thus the gel electrophoretic comigration experiments and the immunoprecipitation results both suggest that t1 is cuticle protein CP1 and t2 is CP2 and that the t3/4 spot is probably CP3 or CP4, or both.

## Genetic Variants of Larval Cuticle Proteins

Several Drosophila melanogaster wild-type strains make cuticle proteins having altered mobilities on native and two-dimensional gels (Fristrom et al., 1978). We have taken advantage of these genetic variants to gain more information about which of the in vitro translation products corresponds to a particular cuticle protein. One strain, denoted 2/3 (Fristrom et al., 1978), shows the following cuticle protein pattern for extracted cuticle proteins on two-dimensional gels (not shown): CP2 and CP3 are missing and replaced by two new proteins denoted CP2v and CP3v. These proteins have estimated molecular

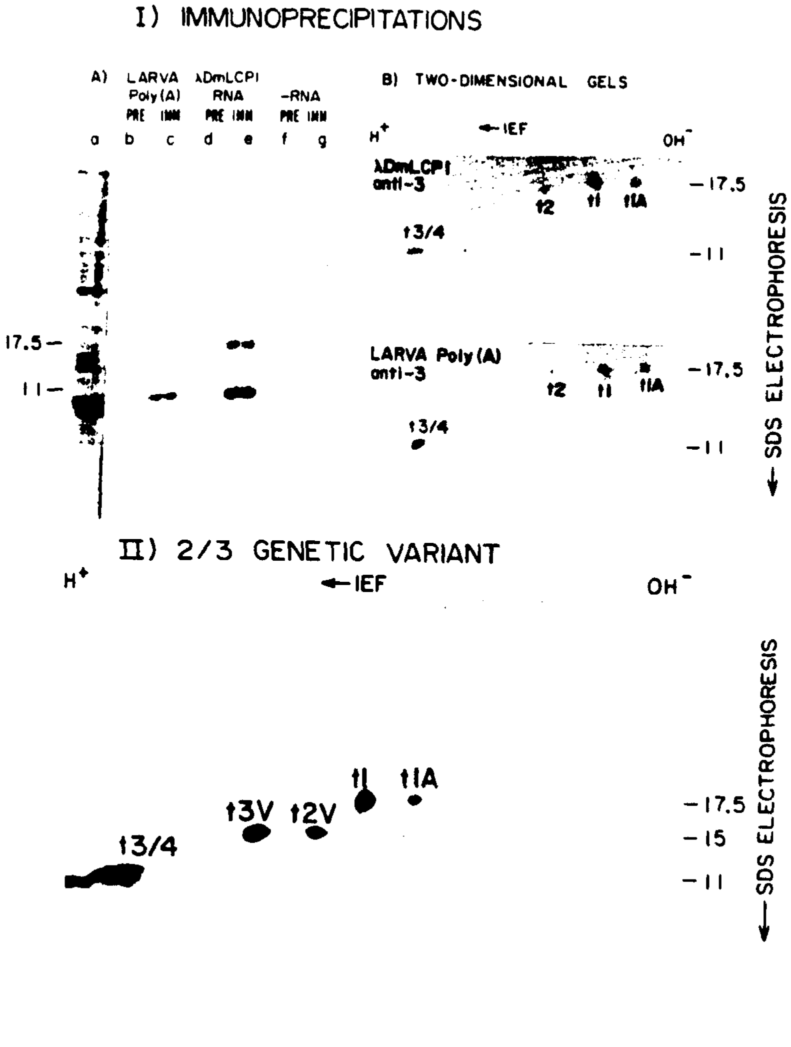


Figure 2. Identification of Translated Polypeptides by Immunoprecipitations and Genetic Analysis

(I) Immunoprecipitations by absorption to Staphylococcus A membranes were performed as described in Experimental Procedures. (A) One-dimensional SDS gel. Antisera were used as follows: lane a, none; lanes b, c, d, e, f and g, anti-CP3. Translation products treated with antisera were as follows: lanes a, b and c, translated products of total poly(A) RNA from late third instar larvae; lanes d and e, translation products from λDmLCP1-selected RNA; lanes f and g, no exogenous RNA added to the in vitro translation system. In alternating lanes are preimmune sera (lanes b, d and f) and sera from immunized rabbits (lanes c, e and g). (B) Portions of two-dimensional gels. λDmLCP1 anti-3: Translated polypeptides of λDmLCP1-selected RNA that bind to anti-CP3 serum. Larva Poly(A) anti-3: Translated products of late larval poly(A) RNA reacted with anti-CP3 serum.

(II) Translation products from the 2/3 cuticle protein variant strain. In vitro translation products from RNA selected by λDmLCP1 with total RNA from late third instar 2/3 larvae of the variant. Selections carried out in parallel with total RNA from Canton S third instar larvae producted the typical pattern of t1A, t1, t2 and t3/4 polypeptides seen in Figure 1B (data not shown).

weights and isoelectric points of 15 kilodaltons and 5.7 for CP2v and 12 kilodaltons and 5.5 for CP3v. CP2v is nonabundant in crude cuticle preparations and absent in highly purified cuticle preparations, suggesting that it is either readily lost in the purification process or resides principally in the underlying epidermal layer.

When λDmLCP1 is used to select RNA isolated from this strain, two-dimensional gel analysis of the translated polypeptides yields the following result (Figure 2, part II): t1A, t1 and t3/4 protein spots are still observed; t2, however, is absent. In addition, two new protein spots, t2v and t3v, are observed. Spot t2v has the same isoelectric point as CP2 but is 0.0–0.5 kilodaltons greater in molecular weight than CP2v; similarly, t3v has the same isoelectric point as CP3v but is 2.5 kilodaltons greater in molecular weight. Since t2 and CP2 are both absent in this strain and replaced by variant proteins, the simplest explanation is that t2 is either the precursor to CP2 or the mature polypeptide. That another new in vivo protein CP3v has an in vitro counterpart, t3v translated from RNA selected by λDmLCP1, suggests that the gene encoding CP3 is on this ċlone. However, the t3/4 spot is still present in the translation products from the 2/3 RNA. We present evidence below that there are two separate genes on λDmLCP1, which are not closely related in sequence, and each of which codes for an in vitro translation product that is part of the t3/4 spot. DNA and protein sequencing studies (M. Snyder, M. Hunkapiller, D. Yuen, D. Silvert, J. Fristrom and N. Davidson, unpublished) show that one of these genes codes for CP3. The other gene therefore probably encodes CP4, which remains unchanged in the 2/3 variant strain.

### Cytological Localization on Polytene Chromosomes

Electrophoretic variants such as the 2/3 strain have been used to genetically map four of the cuticle proteins (Fristrom et al., 1978; C. Chihara, D. Kimbrell and J. Fristrom, unpublished). CP5 has been localized to chromosome three, while CP1, ĊP2 and CP3 have all been mapped to chromosome two in the region of 44–50. CP4 is unmapped. As shown in Figure 3, we have cytologically localized λDmLCP1 on polytene chromosomes by in situ hybridization (Gall and Pardue, 1971). Grains were localized only over region 44D, consistent with the genetic data, and at no other chromosomal sites. In addition, the number of grains exposed at 44D was equal to that of a single-copy standard included in these experiments, which hybridized to 38A (Hirsh and Davidson, 1981). Thus the cloned insert appears to be mainly a single-copy sequence from region 44D.

### The 44D Clones Encode a Gene Cluster

Since four polypeptides are translated from RNAs selected by these clones, we determined the number and the positions of the genes on the cloned inserts by electron microscopic R looping and by Southern blotting. Poly(A) RNA from total late third instar larvae was hybridized to Charon 4 clone DNA under conditions favorable for R-loop formation (Kaback et al., 1979). Unhybridized RNA was removed by gel filtration, and the DNA was spread for electron microscopy. As shown in Figure 4A, multiple genes were observed in λDmLCP1. Three small R loops, each 0.5 ± 0.1 kb in length and thus of a size expected for genes coding for small proteins, are clustered within 4.5 kb of DNA. These are named genes II, III and IV in correlation with the cuticle protein they are shown to encode (see below). Genes II (see Figures 4A and 4B) and III are separated by 0.85 ± 0.1 kb, while genes III and IV are 1.9 ± 0.1 kb apart. Seventy percent of the λDmLCP1 molecules (n = 268) contains all three of these R loops, and another 25% contains two of the three. No intervening sequences were observed in any of these three genes. Another gene, gene V, was observed, although relatively infrequently; only 5–10% of the molecules contains this R loop. Since the DNA for this gene is in excess and hybridizations were carried to a Cot at which all the complementary RNA is driven into R loops (Kaback et al., 1981), the abundance of gene V RNA in late larval poly(A) RNA could be estimated to be 2–4 × $10^{-5}$. In contrast, the abundance of the RNAs complementary to the other three genes is estimated to be about $10^{-3}$, as determined by immunoprecipitations of the translated products for the genes. Gene V is 1.0–1.5 kb in length and is often seen tangled in a fashion consistent with the presence of a small intervening sequence. Because of its low abundance in late third instar larval RNA, we have not characterized this gene further.

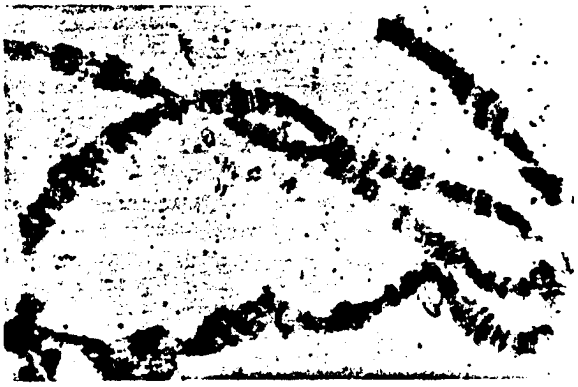


Figure 3. Localization of λDmLCP1 by in Situ Hybridization

R looping to the overlapping clone λDmLCP3 revealed more of the gene cluster (Figure 4C). Genes II and III are seen on the overlapping sequence shared with λDmLCP1 (see below), and a fifth, abundantly expressed gene, I, was found. This gene lies 2.82 ± 0.12 kb from gene II and, like the other abundantly expressed genes, it is 0.5 ± 0.1 kb in length and has no observable intervening sequences.

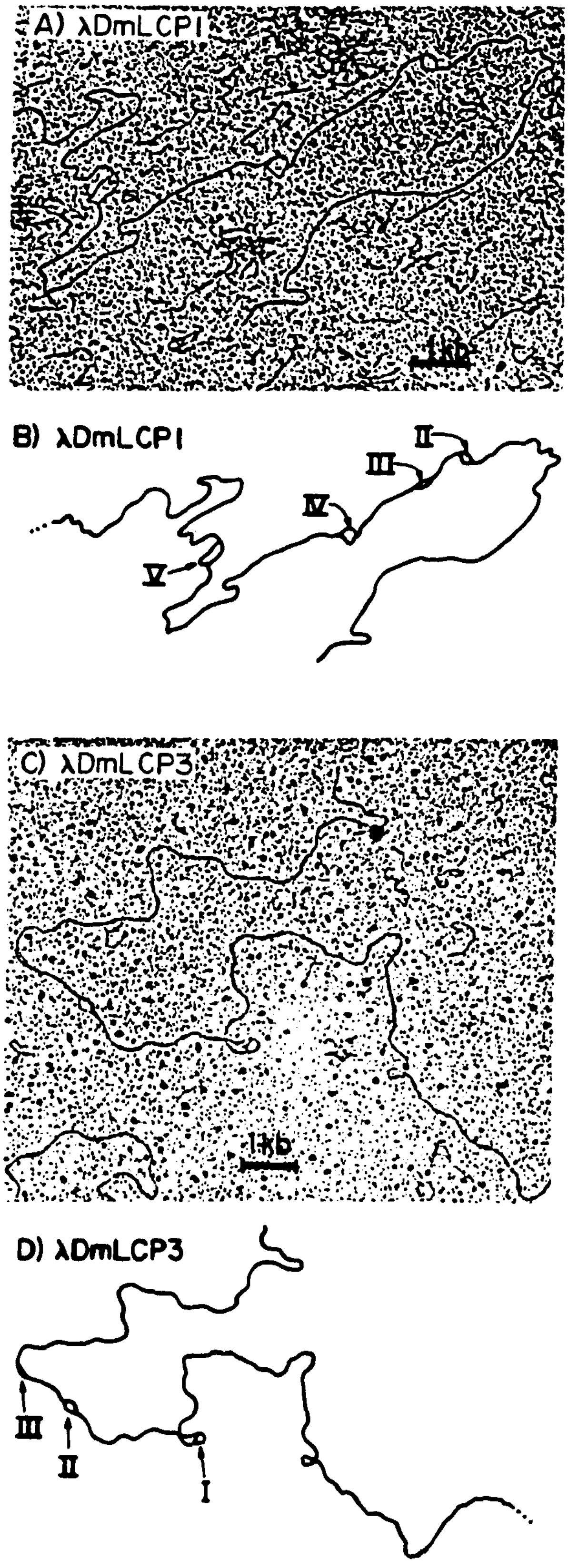


Figure 4. R Loops of 44D Clone DNA to Late Larval Poly(A) RNA

R loops of total late larval poly(A) RNA to whole lambda clone DNA were prepared as described in Experimental Procedures. (A) λDmLCP1; (B) tracing of (A); (C) λDmLCP3; (D) tracing of (C). In each figure, the 10.9 kb right arm of the lambda DNA as well as the cloned insert is shown.

Multiple genes were also established by restriction endonuclease mapping and Southern blotting experiments. Gel blots were performed on restriction endonuclease digests of the phage DNAs and probed with a representative, calf-thymus-DNA-primed cDNA made to total late larval poly(A) RNA. Examples of these blots are shown in Figure 5. Hybridization to the nonabundant gene V can just be detected in these experiments (see Figure 5B, lane Sac), consistent with the abundance measured by R looping. The restriction mapping data, R looping and DNA blot data are summarized in Figure 6.

### Correlating 44D Genes with Cuticle Proteins

Since multiple genes are observed in this region and multiple proteins were translated from RNAs selected by λDmLCP1 and λDmLCP3, we determined which genes, I, II, III and IV, code for which proteins, t1A, t1, t2 and t3/4, using subcloned DNA fragments containing individual genes. The subclone denoted pCPI-11 contains an insert coding for gene I, pCPII-7 for gene II, pCPIII-9 for gene III and pCPIV-8a for gene IV (see Figure 6). The results of the RNA selection and translation experiments with these clones are presented in Figure 7. Gene III and gene IV both select an RNA that translates to give a polypeptide at the t3/4 spot (Figure 7, parts III and IV). Moreover, each of the genes, III and IV, translates approximately equal amounts of t3/4, even under stringent RNA selection conditions. (Filters containing either gene III or gene IV bound RNA were subjected to final washes in 10 mM Tris-HCl, 2 mM EDTA [pH 7.8] at 75°C prior to elution and translation, with less than 50% loss of t3/4 polypeptide in each case [data not shown].)

Each of the genes I and II, selects RNA that translates to give all three 17.5 kilodalton polypeptides, t1A, t1 and t2 (Figure 7, parts I and II). However, the relative amounts of the three polypeptides reproducibly differ in that gene II is slightly more efficient at selecting RNA for t2 and gene I preferentially selects RNAs for t1A and t1. This suggests that t2 is encoded by gene II; t1A or t1 or both are encoded by gene I. As shown previously, t2 is distinct from t1A and t1 at the RNA level because t2 is not found in RNA selected by λDmLCP1 in the variant strain 2/3, while t1A and t1 are found.

### Genomic Representation of λDmLCP1 and λDmLCP3 Sequences

As shown above, genes I and II are closely related in sequence, as indicated by the fact that either one selects the several RNAs that translate to give the same three polypeptides, t1A, t1 and t2. One simple explanation for the three translation products would

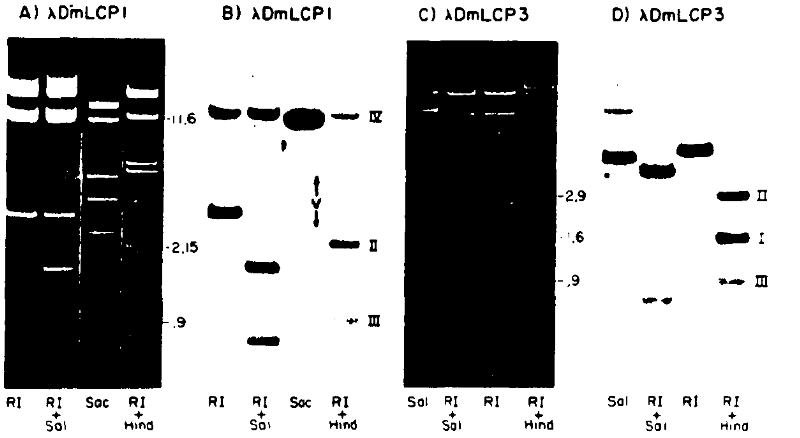


Figure 5. Southern Blotting Analysis of λDmLCP1 and λDmLCP3 Coding Regions

λDmLCP1 and λDmLCP3 DNAs were digested with the restriction endonucleases indicated and separated on a 0.7% agarose gel. (A) Shows ethidium staining of the DNA for λDmLCP1 and (B) for λDmLCP3. Gel blots were prepared and hybridized to representative, calf-thymus-primed cDNA made to total late larval poly(A) RNA. (B and D) Show the pattern of fragments that hybridize. The Sac lane in (B) shows a longer exposure than the other lanes. Units given on the vertical axis are lengths in kb. In the last lane(s) a roman numeral is placed next to a hybridizing fragment to indicate the gene contained on the fragment.

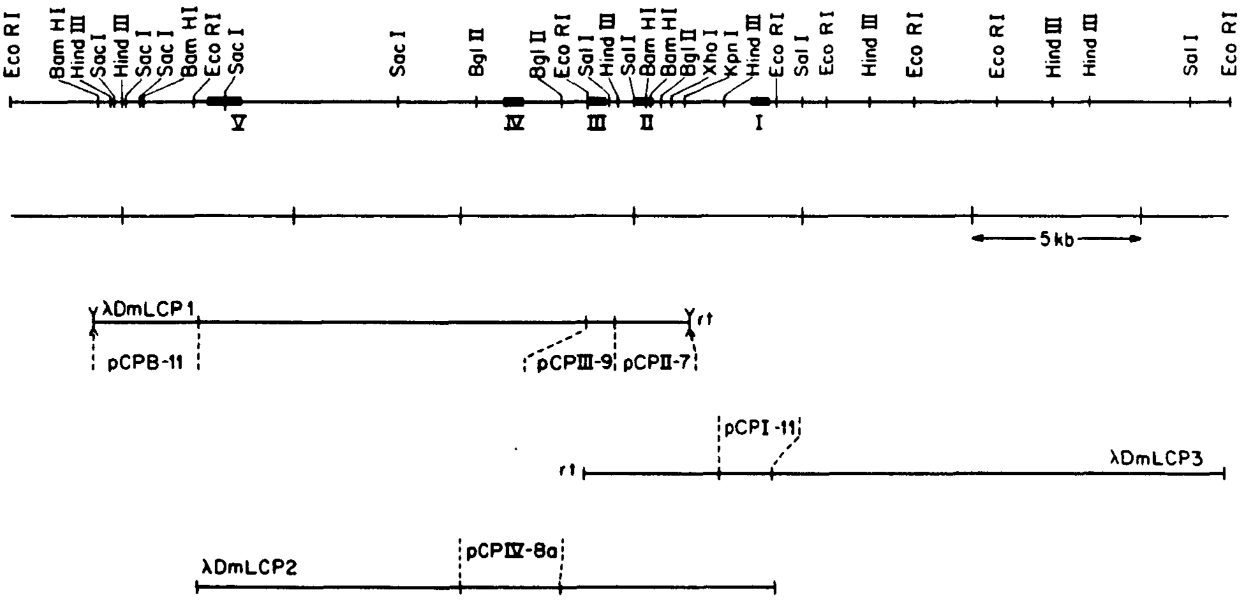


Figure 6. Restriction Endonuclease Map of the 44D Region

The upper diagram shows the composite map of the 44D region as derived from lambda clone inserts depicted below the map. Subclones used in this study were derived as indicated. Note that the λDmLCP1 insert terminates with synthetic Eco RI sites. λDmLCP3 and λDmLCP2 inserts contain natural Eco RI sites at their ends. The restriction sites indicated correspond to all the sites found in this region except that Xho, Bam and Sac are unmapped outside λDmLCP1, Bgl II sites were unmapped outside λDmLCP2 and no restriction sites were mapped in the 2.4 kb segment to the left of λDmLCP1. The length of the leftmost 2.4 kb in this diagram was deduced from lambda phage isolated in the process of chromosomal walking. The notation (rt) next to one end of the clone insert indicates that that end is attached to the 10.9 kb right arm of Charon 4 DNA.

be the existence of another gene closely related in squence to genes I and II but not included on the segment of DNA spanned by the clones λDmLCP1 and λDmLCP3. By immunological criteria, four of the major cuticle proteins share some antigenic determinants; conceivably, these genes could be related in sequence.

We have therefore carried out several DNA blotting experiments to determine whether there are other genes in the genome that are related in sequence to those included within the cluster described here and to study in further detail the sequence relations of the genes within the cluster. Our conclusions, based on the experiments described immediately below, are that genes I and II are closely related in sequence, as expected from the RNA selection experiments, that otherwise the genes and the sequences within the cluster are mainly single-copy and there are no closely related genes outside of the cluster but that there are several short or partial regions of homology between segments within the cluster. In particular, a sequence in gene III shares weak homology with some sequence within a 5.6 kb DNA fragment that contains gene IV (Figure 8). There also may be several other weak homologies, as discussed below.

We first examined the average copy number of the

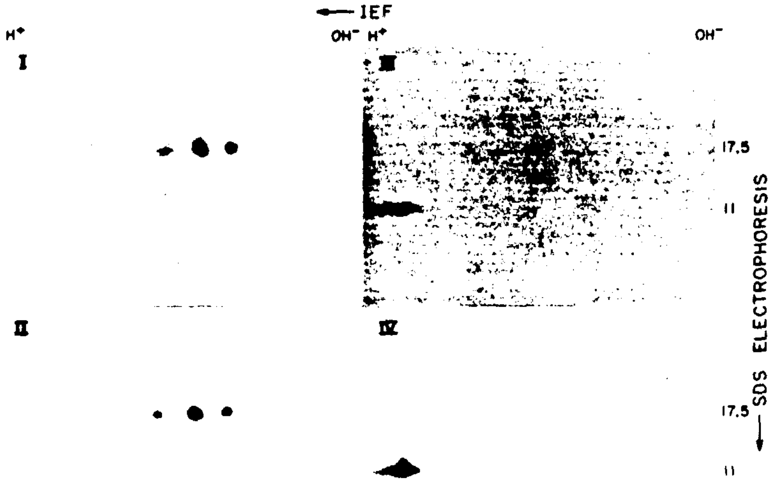


**Figure 7. Translation of RNAs Selected by Individual Genes**

Four subcloned inserts containing individual genes, as indicated in Figure 6, were used to select RNA from total late larval poly(A) RNA. This RNA was translated in vitro and the proteins synthesized were analyzed on two-dimensional gels. (I) Gene I subclone, pCPI-11; (II) gene II subclone, pCPII-7; (III) gene III subclone, pCPIII-9; (IV) gene IV subclone, pCPIV-8a.

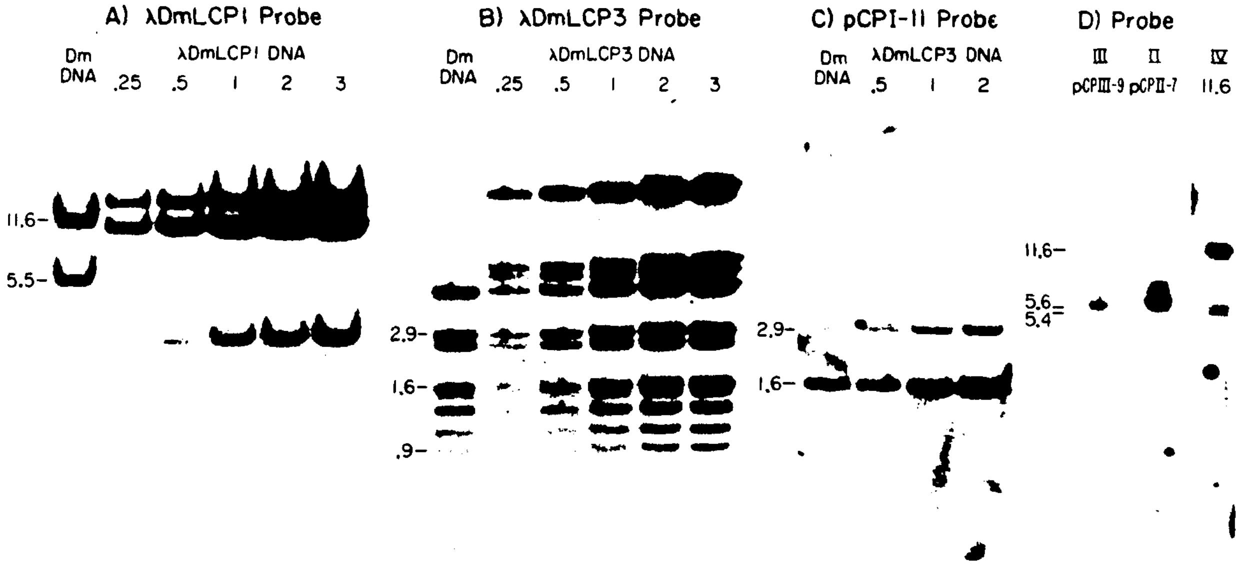


**Figure 8. Genomic Representation of λDmLCP1 and λDmLCP3 Sequences**

Genomic Drosophila pupal DNA (5 μg) was digested with restriction endonucleases and fractionated on a 0.7% agarose gel. (A, B and C) The copy numbers of the hybridizing Drosophila genomic fragments were estimated by comparison with the adjacent reconstruction lanes containing digested calf thymus DNA plus amounts of lambda clone DNA equivalent to the copies per haploid genome indicated in the figure. Gel blots were prepared and probed as follows. (A) The DNAs were digested with Eco RI. The adjacent lanes contain λDmLCP1 DNA and the probe was $^{32}$P-labeled nick-translated λDmLCP1; (B) the DNAs were digested by Eco RI plus Hind III. The adjacent lanes contain λDmLCP3 DNA and the probe was $^{32}$P-labeled nick-translated λDmLCP3; (C) the DNAs were digested with Eco RI plus Hind III. The adjacent lanes contain λDmLCP3 DNA and the probe was $^{32}$P-labeled pCPI-11 DNA. In the reconstruction lanes of (A), the right arm of the lambda vector was not resolved from the 11.6 kb insert fragment. Furthermore, there are two 3.05 insert fragments generated by digestion; these correspond to the two ends of the insert and each contains one artificial Eco RI site. They correspond to the genomic 5.6 and 5.4 kb fragments, which were also not resolved in this gel. In both (A) and (B), the reconstruction lanes contain hybridizing bands caused by the vector arms, which do not match any bands in the genomic lanes. (D) Shows an Eco RI digest of Drosophila pupal DNA probed with nick-translated probes of pCPIII-9; pCPII-7; and a gel-isolated 11.6 kb Eco RI fragment from λDmLCP1 that contains genes IV and V. In these gels, the 5.4 kb and 5.6 kb natural Eco RI fragments were resolved. Units indicated in the figure are in kb.

entire cloned DNA segments of λDmLCP1 and λDmLCP3 in the Drosophila genome. In each case, gel blots of pupal DNA were hybridized to nick-translated probes from the cloned bacteriophage DNA and compared to blots of carrier DNA with amounts of cloned DNA equivalent to 0.25, 0.5, 1, 2 and 3 copies per haploid Drosophila genome. As shown in Figures 8A and 8B, the DNA fragments of genomic DNA that hybridized have the same length as the natural DNA fragments of the cloned insert (see Figure 6). The intensities of hybridization indicate that the majority of the sequences in the cloned DNAs are present as one copy within the Drosophila genome. However, upon long exposures, using the λDmLCP1 probe, we detected three additional very faint bands (data not shown). These bands probably arise from a short sequence therein that is repeated several times in the genome.

Additional experiments were performed with subcloned fragments. Genomic DNA was subjected to double digestion with Eco RI and Hind III, blotted and hybridized to the subcloned gene I probe, pCPI-11 (see Figure 6). The hybridization intensities were compared to those of digests of 0.5, 1 and 2 copies of λDmLCP3 as shown in Figure 8C. The only genome bands that hybridize have lengths of 1.6 and 2.9 kb, and these have single-copy intensities. These are the expected lengths for the cloned segments containing genes I and II. Thus the experiment confirms the sequence relatedness of these two genes, indicates that each is single-copy in the genome and shows that there are no other cross-hybridizing sequences within the genome.

These conclusions were further tested with the subcloned fragments of gene II. The subcloned fragments are the 400 bp Sal I–Bam HI DNA fragment and the adjacent 120 bp Bam HI DNA fragment (see Figure 6). Together, these subclones contain 90–100% of the protein coding region of gene II and the 5′-untranslated region, as determined by DNA sequencing results to be reported later. When Eco RI–digested genomic DNA was probed with a mixture of these two $^{32}$P-labeled nick-translated subclones, a 5.6 kb band hybridized at the intensity that would be expected if it were the 5.6 kb fragment of λDmLCP3 containing genes I and II (data not shown). No other bands were observed.

Additional studies were carried out with subcloned probe pCPII-7 for gene II and pCPIII-9 for gene III, and with the gel-isolated 11.6 kb fragment from λDmLCP1 containing gene IV and gene V. These probes were hybridized to gel blots of Eco RI–digested genomic DNA (Figure 8D). The gene II probe hybridizes only to the expected 5.6 kb fragment. This fragment contains both genes I and II, which, as shown above, cross hybridize strongly. The pCPIII-9 gene III probe hybridizes predominantly to the same 5.6 kb fragment as expected (see Figure 6). Weak hybridization is observed to an 11.6 kb fragment.

The nature of this weak hybridization was further studied by probing a blot of a Sac I–Eco RI–Hind III triple digest of λDmLCP1 with pCPIII-9. As expected, there is strong hybridization to the 0.9 kb band of origin. In addition, there is weak hybridization to the 5.6 kb band containing gene IV (2–4% relative to the strong band; data not shown). Thus there is weak homology between some sequences in pCPIII-9 and some sequence within the gene-IV-containing band. However, genes III and IV do not cross hybridize strongly.

In an Eco RI genome blot with the 11.6 kb Eco RI segment of λDmLCP1 as probe (Figure 6), there is the expected strong hybridization to the band of origin at 11.6 kb. In addition, there is weak hybridization to a 5.6 kb Eco RI fragment. This is presumably the fragment containing genes III, II and I, and the hybridization must be partly due to the weak homology discussed above. In addition, there is weak hybridization to a 5.4 kb Eco RI genomic fragment. We believe this is the 5.4 kb fragment at the extreme left of the map in Figure 6. This interpretation is supported by the observation that the left 3.0 kb fragment of λDmLCP1, pCPB-11 (with one artificial Eco RI end), hybridizes weakly to the 11.6 kb internal Eco RI fragment of λDmLCP1 (data not shown).

### The Abundant Genes of the 44D Cluster Are All Expressed in the Same Tissue and Stage of Development

The data presented above demonstrate that a family of genes is clustered. To determine how the genes in this cluster are expressed relative to each other in terms of stage of development and tissue type, we performed the following experiments. Poly(A) RNAs from a number of stages of Drosophila development were isolated, subjected to electrophoresis in the presence of methylmercury hydroxide, transferred to diazotized paper and probed with a nick-translated probe synthesized from whole clone λDmLCP1 DNA. Since genes I and II are similar in sequence, this probe is expected to hybridize to all RNA species from the genes in the cluster. The result is shown in Figure 9A. As is seen in lane 4, these RNAs are abundantly expressed in late larval RNA, and all migrate in one broad band of 500–600 bp, consistent with the R-looping results. None of these messages can be detected in early or late embryos, in pupae at the time of pupation, in 70 hr pupae or in adults. The limit of detection in this experiment (with longer exposures) is estimated to be 2% of the amount expressed in late larvae. Low levels of these messages can be found in second instar larvae and collectively are about 10% as abundant as in late third instar larval RNA (compare the second instar RNA lane with the lane containing one tenth the amount of third instar RNA). In addition, the messages for these genes are much more prevalent in poly(A) RNA from a larval integument preparation that is enriched for epidermal cells actively synthesizing cuticle proteins.

That all the abundant genes are expressed in both larval poly(A) RNA and even more so in larval integument RNA was shown as follows. The λDmLCP1 DNA was digested with the three enzymes, Eco RI, Hind III and Sac I, which separates genes II, III and IV on

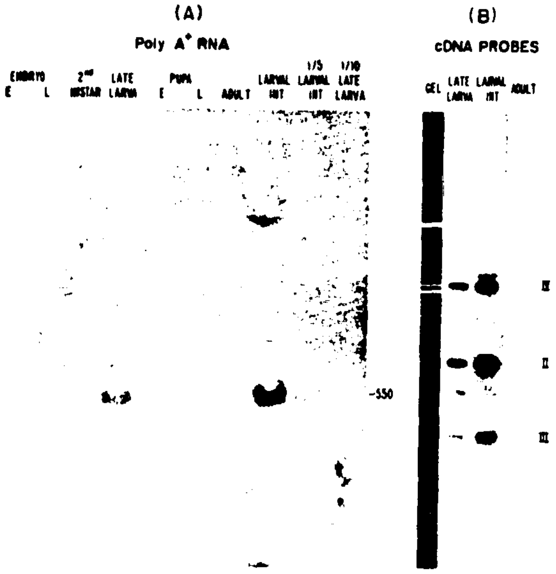


Figure 9. Developmental Expression of the λDmLCP1 Encoded Genes

(A) Poly(A) RNA was isolated from several stages of Drosophila development and 1 μg of each RNA was subjected to electrophoresis in a 1% agarose gel containing 10 mM methylmercury hydroxide. RNA blots were prepared and hybridized to labeled nick-translated λDmLCP1 clone DNA. Starting from the left lane, RNA was isolated from early embryos 0–4 hr after egg laying; late embryos 16–20 hr after egg laying; second instar larvae; late third instar larvae; early pupae 33 hr after late larval collection; later pupae 73 hr after late larval collection; adults 0–24 hr after eclosion; late third instar integuments. In the last two lanes, lower amounts of RNA were used. The penultimate lane contains 0.2 μg (one fifth the previous amounts) late third instar integument RNA. The last lane contains 0.1 μg late third instar RNA. (B) λDmLCP1 DNA was digested with the restriction endonucleases Eco RI, Hind III and Sac I. This digestion separates genes II, III and IV into restriction fragments of different lengths. Fragments were separated by electrophoresis in one lane of a 0.7% agarose gel and blotted onto nitrocellulose. The blot was cut into strips, and each strip was hybridized in parallel to a different calf-thymus-primed cDNA probe. The lanes are: (gel) ethidium stain of digested fragments; (late larva) hybridization to total late larval cDNA probe; (larval int.) hybridization to late larval integument cDNA probe; (adult) hybridization to adult cDNA probe. The band larger than IV, which weakly hybridizes, is caused by an incompletely digested DNA fragment.

restriction fragments of different sizes. DNA blots were hybridized to cDNA probes synthesized to total late larval poly(A) RNA and larval integument poly(A) RNA in parallel. As shown in Figure 9B, genes II, III and IV, are all expressed in both RNA populations, and each is enriched in the larval integument by the same relative amount of about 4-fold. Thus these genes are actively expressed in the same developmental stage and tissue. Gene I, because of its close homology with gene II, is expected to show the same pattern of expression. At no point in these experiments could we detect any gene V hybridization. Because of its greater length, RNA complementary to gene V should have been well separated from gene II, gene III and gene IV, RNAs in the RNA gel blot.

## Discussion

The data presented above show that there is a cluster of genes at region 44D of the Drosophila genome. There are four closely spaced genes (I, II, III and IV, Figure 6), which are abundantly expressed in the same tissue and at the same time during development. Gene V, which is adjacent to the cluster, is nonabundantly expressed in late larval poly(A) RNA and has not been studied further.

Genes I and II, which cross hybridize and are otherwise unique in the genome, select RNAs that translate to give the peptides t1A, t1 and t2. These three peptides have the same molecular weight. They differ in isoelectric points by small increments in pH such that they probably differ by only one or two residues in content of charged amino acids. The relative amounts of the translation products from the RNAs selected by genes I and II suggest that gene II codes for t2, whereas gene I codes for t1 and t1A. Immunoprecipitation studies show that t1A, t1 and t2 are all cuticle polypeptides. Gel electrophoretic comigration studies and studies of the translation products from the RNA of the 2/3 variant indicate that t2 is CP2 or a precursor thereof. Gel electrophoretic comigration comparison indicates that t1 is CP1. This leaves the identity of t1A undetermined. We presume that t1A and t1, as products of gene I, are related by posttranslational modification. Protein sequencing of CP1 and CP2 and DNA sequencing of genes I and II have confirmed our assignments; CP1 is encoded by gene I, and CP2 by gene II (M. Snyder, M. Hunkapiller, D. Yuen, D. Silvert, J. Fristrom and N. Davidson, unpublished).

The translation products of genes III and IV are found in the t3/4 spot. These migration experiments are not decisive as to whether t3/4 contains precursors to mature polypeptides for CP3, CP4 or CP5. Since CP5 does not map at region 44D and is not immunoprecipitated with antiserum to CP3, t3/4 does not include CP5 polypeptides. Translation of the RNA from the 2/3 variant suggests that one component of t3/4 is CP3, for which there is a counterpart in the variant. Decisive evidence that gene III codes for CP3 has been obtained by protein and DNA sequence studies (M. Snyder, M. Hunkapiller, D. Yuen, D. Silvert, J. Fristrom and N. Davidson, unpublished). Since gene IV also codes for a spot at t3/4, it probably encodes CP4, but this identification is not decisive.

These sequence studies indicate that t3/4 is about 20 amino acids greater in molecular weight than CP3 because of a hydrophobic signal peptide at the amino terminus. The gel in Figure 1B shows that t3/4 does have a slightly greater molecular weight than CP3. CP1 and CP2 are secreted proteins and thus would also be expected to have signal peptides at their amino termini. The gels in Figure 1 do not show any difference in molecular weight between t1 and CP1 or between t2 and CP2. However, a difference of about 10 amino acids or less may not have been detected.

D. Silvert and J. Fristrom (in preparation) have shown that CP1, CP2, CP3 and CP4 share common antigenic determinants. Our preliminary sequence data indicate that CP1 and CP2 have a high degree of amino acid sequence homology, whereas CP2 and CP3 are more distantly related (M. Snyder, M. Hunkapiller, D. Yuen, D. Silvert, J. Fristrom and N. Davidson, unpublished). These results are consistent with the observation that genes I and II cross hybridize, but genes II and III do not at our hybridization criteria (which allow hybridization with about 20% or less mismatch). In addition, genes III and IV are not closely related in sequence. Thus the several genes of this cluster probably evolved by duplication of a primordial cuticle protein gene and then diverged; genes I and II were probably formed in the most recent duplication.

In Drosophila, gene clusters whose members are coordinately expressed in the same tissue and at the same time of development have now been found in the case of the chorion genes (Spradling et al., 1980; Griffin-Shea et al., 1980), two yolk protein genes (Barnett et al., 1980), some heat-shock genes (Corces et al., 1980) and the 68C salivary gland puff genes (E. Meyerowitz and D. Hogness, personal communication). The larval cuticle gene cluster studied here is another example. The clustering of genes may be an essential part of a general mechanism of activating a segment of the chromosome for transcription and thereby controlling the coordinate expression of genes. It is interesting in this regard that the larval cuticle gene cluster contains a nonabundantly expressed gene, gene V, which is separated from the closely spaced and abundantly expressed genes by 8 kb. We therefore speculate that if a general mechanism for activating a region of DNA for transcription does exist, genes that reside toward the periphery of such a region may be less accessible for transcription and hence expressed at a much lower level. Thus the positioning of gene V could be significant in determining its level of expression.

The cuticle gene cluster is also interesting in that it is directly or indirectly induced by ecdysone.

## Experimental Procedures

### Materials

The restriction endonucleases Eco RI, Pst I and Hind III were prepared by M. Allonso according to standard procedures. All other restriction enzymes and T4 DNA ligase were purchased from New England BioLabs. E. coli DNA polymerase and RNA polymerase were obtained from Boehringer-Mannheim. AMV reverse transcriptase was a gift from J. Beard. Staphylococcus A ghosts were a gift from J. Frelinger. We purchased $\alpha$-$^{32}$P-labeled triphosphates, 410-800 Ci/mmole, from either New England Nuclear or Amersham; $\alpha$-3,4,5$^{3}$H(N)-leucine, 110 Ci/mmole, was obtained from New England Nuclear.

### Screening the Library

#### *Selecting a Cuticle Clone*

Two Canton S Drosophila (Dm) recombinant DNA libraries in Charon 4 were used; a random shear library prepared by J. Lauer from embryo DNA, with inserts of 12-20 kb, terminated by synthetic Eco RI linkers (Maniatis et al., 1978); and the Eco RI partial digest libary of Dm pupal DNA with inserts of 12-20 kb described by Yen et al. (1979) and Davidson et al. (1980).

The screening for a recombinant phage with an insert coding for larval cuticle genes was accomplished in four steps. The first two steps are described in detail in J. Hirsh and N. Davidson (submitted). Clones coding for abundantly expressed mRNAs in third instar larvae were first selected by screening the random shear libary using an oligo(dT)-primed cDNA probe made to poly(A) RNA extracted from the integument of late third instar larvae.

These phages were then tested by differential plaque hybridization. Phage clones that gave strong positive hybridization signals with oligo(dT)-primed cDNA probes to larval integument poly(A) RNA as compared to probes for embryo poly(A) RNA were chosen. Of the 37 phages selected, DNA was isolated from 16: 1 $\mu$g of each DNA was denatured in 10 $\mu$l of 0.1 M NaOH, 1 NaCl, 1 mM EDTA, neutralized by the addition of 1 $\mu$l 10 N HCl plus 10 $\mu$l of 0.5 M Tris-HCl (pH 7.4), 1 M NaCl at 0°C and absorbed onto 0.45 $\mu$ nitrocellulose filters (Schleicher and Schuell). DNA samples were spotted in duplicate on each of four filters, which were then washed in 2× SSC (SSC = 0.15 M NaCl, 0.015 M Na citrate) and baked. Each filter was then hybridized to a cDNA probe made to poly(A) RNA from whole 16-hr embryos; late third instar larvae; pupae 33 hr past pupariation; or late third instar larval integument fraction prepared as described by J. Hirsh and N. Davidson (submitted). The four phages that showed the highest hybridization signal to the larval integument probe relative to the three others were chosen. These were tested in the last selection step, which involved hybridization selection of RNA, in vitro translation and two-dimensional gel analysis of polypeptides as described in Results.

#### *Chromosomal Walking*

Probes were synthesized from the subclones pCPII-7, pCPIII-9 and pCPB-11 (see Figure 6) representing the two ends of the λDmLCP1 insert. The DNAs were individually labeled by nick translation, and $10^{7}$ cpm from each were mixed to obtain sequences flanking the λDmLCP1 insert on both sides. This probe was used to screen 60,000 phages from the Eco RI partial digest library by in situ plaque hybridization (Benton and Davis, 1977) at a density of 5,000 phages per 90 cm plate. Hybridization conditions were as described by Mullins et al. (1981). Filters were washed as described below. After screening, the total number of hybridizing phages was 29, which is about the number expected for single-copy probes. DNA was prepared from ten of these positive phages. Both by restriction endonuclease mapping and by probing gel blots with nick-translated subcloned probes, all ten phages were determined to be contiguous with λDmLCP1 sequences.

### Nucleic Acid Preparations

#### *Recombinant DNAs*

Charon phages were grown on plates as described in Yen and Davidson (1980). Following lysis, phages were extracted from top agar, treated with 1 $\mu$g/ml each of RNAse A and DNAase I, banded in a cesium chloride step gradient and rebanded two times in cesium chloride density gradients. Phage DNA was extracted and stored at 4°C in 0.01 M NaCl, 0.01 M Tris-HCl (pH 7.8), 1 mM EDTA.

Plasmid DNA was prepared by centrifuging bacterial lysates from chloramphenicol-treated cells in ethidium bromide, cesium chloride density gradients as described in Fyrberg et al. (1980). High molecular weight Drosophila pupal DNA was isolated by N. D. Hershey as described in Fyrberg et al. (1980).

#### *Undegraded Cellular RNA*

Undegraded total cellular RNA was prepared by homogenizing whole animals in 4 M guanididium thiocyanate, 1 M $\beta$-mercaptoethanol, 0.05 M sodium acetate, 0.001 M EDTA (pH 6) and banding in cesium chloride as described in Fyrberg et al. (1980). Pupae were staged by taking crawling, late third instar larvae and incubating them for 33 or 73 hr. Adults were collected 0-24 hr after eclosion. Six hours after the larvae had been collected, pupariation had ensued as judged by anterior spiracle formation in more than 95% of the animals. For the preparation of larval integument RNA, late third instar larvae were squashed twice on a glass plate with a rolling pin to remove the internal viscera. The outer integuments were quickly gathered and, within 30 sec after squashing, were transferred to the 4 M guanidium thiocyanate solution and homogenized. The yield of total RNA from the integument fraction was one tenth to one twentieth that of total late third instar larval RNA (0.5–1 μg RNA per integument).

Poly(A) RNA was selected by oligo(dT) cellulose (Collaborative Research, Inc.) chromatography (Bantle et al., 1976). RNA was bound to oligo(dT) in 0.5 M NaCl, 0.01 M Tris-HCl (pH 7.4), 1 mM EDTA, 0.1% SDS, washed in 0.2 M NaCl, 0.01 M Tris-HCl (pH 7.4), 1 mM EDTA and eluted in 10 mM Tris-HCl (pH 7.4), 1 mM EDTA. The poly(A) RNA was recovered by ethanol precipitation, redissolved in double-distilled water and stored at −70°C. Binding to oligo(dT) once results in 50% of the eluted material containing poly(A) tracts (Bantle et al., 1976; and unpublished data). Poly(A) RNAs were subsequently judged to be undegraded by two means. First, agarose gel electrophoresis in the presence of methylmercury hydroxide showed the characteristic ribosomal RNA bands present in the preparation. Second, 300 ng of the RNA demonstrated a high translation efficiency in a rabbit-reticulocyte-cell-free translation system (see below).

### Restriction Endonuclease Mapping and Gel Electrophoresis of DNA

Restriction endonuclease maps of cloned DNA were generated with single, double and partial digests with various restriction endonucleases. Care was taken to look for fragments larger than 200 bp for λDmLCP1 and larger than 250 bp for λDmLCP3. A DNA fragment 125 bp in length in λDmLCP1 was not missed.

Gel electrophoresis of DNA fragments was carried out with 0.7% agarose gels for examining larger DNA fragments and 5–7.5% acrylamide gels for <0.8 kb DNA fragments. Conditions for electrophoresis of DNA and isolation of DNA fragments by hydroxyapatite are described in Hershey and Davidson (1980) and Fyrberg et al. (1980), respectively. DNA fragments subjected to electrophoresis were transferred to 0.22 μ nitrocellulose (Millipore) according to the procedure of Southern.

### DNA Labeling and Hybridizations

$^{32}$P-labeled cDNA probes were prepared either with oligo(dT) from Collaborative Research, Inc., according to the procedure of Efstratiadis et al. (1975), or with calf thymus (prepared by J. Casey) DNA primers, according to the procedure of Taylor et al. (1976). Only labeled CTP was used.

$^{32}$P-labeled nick-translated probes were synthesized according to standard protocols (Schachat and Hogness, 1973; Maniatis et al., 1975). Either labeled CTP or ATP and TTP was used.

Hybridizations of $^{32}$P-labeled probes to filter bound DNA were performed as follows: In initial experiments, filters were prewashed in 3× SSC, 0.1% SDS and 10× Denhardt's solution for 30 min at 65°C. This step was omitted in more recent experiments. Prehybridization was carried out in 1 M NaCl, 50 mM Tris-HCl (pH 8.3), 0.1% (w/v) SDS, 1 mM EDTA, 10× Denhardt's solution, 10 μg/ml heat-denatured E. coli DNA and 10 μg/ml poly(rA) for 70 min to 2 hr at 67°C with 0.1 ml solution/cm$^2$ of filter. Hybridizations were performed in fresh solution with $^{32}$P-labeled probes. Filter-driven reactions were incubated for 40–48 hr at 67°C. Probe-driven reactions were incubated 16–24 hr at 67°C. Following hybridization, filters were washed for two 15 min periods in 3× SSC, 10× Denhardt's solution, 0.1% (w/v) SDS, 0.1% (w/v) sodium pyrophosphate (NaPP) at 65°C, and then 3–6 times in a similar solution without Denhardt's solution. Filters were subsequently washed in 6–8 changes of 0.1 M NaCl, 50 mM Tris-HCl (pH 8.3), 1 mM EDTA, 0.1% SDS, 0.1% NaPP at 65°C. The filters in Figure 5 received an additional wash in 0.2× SSC for 20 min at 65°C. Filters were blotted dry and exposed to Kodak XR-5 film with an intensifying screen.

### Electrophoresis of RNA and RNA Blots

RNA was subjected electrophoresis in 1% agarose gels containing 10 mM methylmercury hydroxide (Ventron, Inc.) according to the procedure Bailey and Davidson (1976) as modified by Rozek and Timberlake (1979). Electrophoresis was performed at 4 V/cm for 5 hr. RNA was detected by treating the gel with NH$_4$Ac and staining in ethidium bromide.

For RNA blotting experiments, gels were treated, following electrophoresis, for transfer to diazotized paper according to the procedure of Alwine et al. (1978). Preparation and treatment of the diazotized paper was performed according to procedures developed bv B. Seed (personal communication). Following transfer, filters were prehybridized and then hybridized to nick-translated λDmLCP1 probes for 62 hr in 10 ml of solution at 42°C, both according to the procedure of Alwine et al. (1979). Initial washes were as described above; additional washes were performed in eight changes of 0.1 M NaCl, 50 mM Tris-HCl (pH 8.0), 1 mM EDTA, 0.1% SDS, 0.1% NaPP at 65°C. The sizes of the RNA species observed were determined relative to DNA standards subjected to electrophoresis in adjacent lanes.

### Subcloning Fragments of Genomic Clones

DNA fragments (see Figure 8) were subcloned by the procedure of Yen and Davidson (1980). DNA fragments generated by digestion of Eco RI, Eco RI plus Hind III, Bam HI or Bam HI plus Sal I were ligated into the plasmid vector, pBR322 (Bolivar et al., 1977). DNA fragments generated by digestion of Bgl II were subcloned into pKC7 (Rao and Rogers, 1979), a derivative of pBR322.

### Positive Selection and Translation of RNA

RNA was selected by a procedure similar to that described by Ricciardi et al. (1979). Either 10 μg of bacteriophage DNA or 5 μg of plasmid DNA (previously nicked by ultraviolet irradiation) were denatured by heating and applied to nitrocellulose filters in 10× SSC. A prehybridization step was included with 100 μl of 70% formamide, 0.4 M NaCl, 0.1 M PIPES (pH 6.5) for 70 min at 50°C. Hybridizations were then carried out for 3–4 hr at 50°C in 50 μl of solution of the same buffer containing 5 μg poly(A) RNA or 250 μg of total cellular RNA isolated from whole late third instar larvae. Following hybridization, filters were washed ten times in 1 ml of 1× SSC, 0.5% SDS at 65°C then three times in 1 μl of 0.01 M Tris-HCl, 1 mM EDTA (pH 7.8) at 65°C. RNA was eluted in boiling distilled water for 60 sec and recovered by ethanol precipitation. Translation was in a commercial (New England Nuclear) rabbit reticulocyte in vitro translation system. For each translation with 10 μl of reticulocyte lysate, 50 μCi of 110 Ci/mmole $^3$H-leucine was used. Translation was stopped by the addition of 50 ng each of RNAase A and DNAase I at 0°C for 30 min.

### In Situ Hybridizations

In situ hybridizations to salivary gland chromosomes from giant third instar larvae were carried out according to the procedure of Gall and Pardue (1971), with cRNA synthesized according to the procedure of Wensink et al. (1974). In some hybridizations, cRNA made to a unique cloned Drosophila DNA sequence that hybridizes to 38A at the base of 2L (J. Hirsh and N. Davidson, submitted) was added to provide a single-copy reference site on the chromosome.

### R Looping and Electron Microscopy

Hybridization of total poly(A) RNA from whole late third instar larvae to genomic clone DNA was carried out according to the procedures to Kaback et al. (1979). Larval poly(A) RNA (5 μg) was hybridized to 100 ng of genomic clone DNA in 20 μl of 70% formamide, 0.5 M NaCl, 0.1 M PIPES (pH 7.2), 0.01 M EDTA. To ensure maximum complementarity between the RNA/DNA duplexes formed, the temperature was gradually decreased in 2.5°C intervals, starting from 55.0°C (where very few [2%] R loops were formed) to 47.5°C, with each interval lasting 6 ± 2 hr. R loops were separated from unhybridized RNA by gel filtration and spread for electron microscopy. Double-stranded $\phi\chi$174 DNA was included as a size standard.

### Isolation of Larval Cuticles and Cuticle Proteins

Late third instar cuticles were purified according to conditions described in Fristrom et al. (1978). Drosophila late third instar larvae were ground with a Waring blender in Drosophila Ringer's solution (Ephrussi and Beadle, 1939) saturated with phenylthiourea. Cuticles were collected on a nylon screen, reground and washed as described. Cuticles appeared devoid of cellular debris as viewed by light microscopy.

Cuticle proteins were prepared by either extraction in 7 M urea on ice or extraction in O'Farrell lysis buffer containing 0.1% SDS. Undissolved material was removed by centrifugation.

### Electrophoresis of Protein

Proteins were analyzed on one-dimensional SDS 16% polyacrylamide gels run according to the procedure of Laemmli. Either one tenth (3 μl) of the total translation mixture described above or 3-6 μg of total cuticle proteins were analyzed relative to protein standards. Low molecular weight protein standards used were bovine pancreatic trypsin inhibitor, lysozyme, cytochrome c from horse heart, soybean trypsin inhibitor and carbonic anhydrase. To determine the position of migration of cuticle proteins on 16% polyacrylamide gels containing SDS, protein bands were excised from a nondenaturing gel (Fristrom et al., 1978), incubated twice in 1 ml 0.1 M Tris-HCl (pH 6.8) for 20 min each, then in 1 ml Laemmli sample buffer for 20 min and subjected again to electrophoresis on 16% polyacrylamide gels containing SDS.

Alternatively, samples were analyzed on two-dimensional gels as described by O'Farrell (1975). The lysis buffer was modified to contain 0.1% SDS (Brandhorst, 1976). One third of a translation mixture was lyophilized and resuspended in either 20 μl lysis buffer or 20 μl lysis buffer containing cuticle proteins (5-10 μg) for mixing experiments. Cuticle proteins were analyzed with 5-20 μg of larval cuticle proteins. After the first dimension of isoelectric focusing, improved resolution was obtained by fixing proteins in acetic acid and ethanol by a procedure similar to that described by Jackle (1979). First dimensions were then equilibrated and loaded onto 16% acrylamide gels for the second dimension. Molecular weights were determined by subjecting protein standards to electrophoresis in an adjacent slot. Protein bands or spots were visualized by Coomassie staining. Each major cuticle protein spot on a two-dimensional gel was correlated with cuticle proteins CP1 through CP5 by molecular weight determinations, by analysis of cuticle proteins from Drosophila strains that make variant cuticle proteins and by analysis of purified CP1, CP2 and CP3 individually on two-dimensional gels. The resolution of proteins on two-dimensional gels was not improved by changing the range of ampholine buffer from 5-7 to 4-6. Fluorography was performed with Enhance (New England Nuclear), and dried gels were exposed to preflashed Kodak XR-5 X-ray film.

### Immunoprecipitations

Immunoprecipitations were performed by the Staphlococcus A absorption procedures of Kessler (1975). Translated products (10 μl) were incubated with 1 μl of antisera in 20 μl of PBS (150 mM NaCl, 50 mM $NaPO_4$ [pH 7.2]) for 45 min at room temperature. Antigen-antibody complexes were absorbed with 35 μl of 10% (w/v) Staphylococcus A ghosts and incubation for 40 min at 0°C. Cells were pelleted, washed twice with 1 ml of PBS plus 1% (w/v) BSA, then with 1 ml PBS (without BSA). The bound proteins were recovered by adding 25 μl of either Laemmli sample buffer and heating at 70°C for two min or O'Farrell lysis buffer.

### Biosafety

This research was carried out in accordance with National Institutes of Health guidelines.

### Acknowledgments

We are grateful to Dr. J. Fristrom for valuable discussions about the cuticle proteins and their genes, and to D. Silvert and J. Fristrom for a generous gift of antisera. Drs. H. K. Mitchell and E. B. Lewis have advised us in several respects. We are indebted to N. D. Hershey, J. L. Campbell, E. A. Fyrberg and E. Meyerowitz for reading and critizing the manuscript. This research was supported by National Institutes of Health research grants to N. D., and NIH fellowship to J. H. and by training grant support for M. S.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 18, 1981; revised April 23, 1981

### References

Alwine, J. C., Kemp, D. J., Parker, B. A., Reiser, J., Renart, J., Stark, G. R. and Wahl, G. M. (1979). Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper. Meth. Enzymol. *68*, 220-242.

Bailey, J. M. and Davidson, N. (1976). Methylmercury as a reversible denaturing agent for agarose gel electrophoresis. Anal. Biochem. *70*, 75-85.

Bantle, J. A., Maxwell, I. H. and Hahn, W. E. (1976). Specificity of oligo(dT)-cellulose chromatography in the isolation of polyadenylated RNA. Anal. Biochem. *72*, 413-427.

Barnett, T., Pachl, C., Gergen, J. P. and Wensink, P. C. (1980). The isolation and characterization of Drosophila yolk protein genes. Cell *21*, 729-738.

Benton, W. D. and Davis, R. W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science *196*, 180-182.

Bolivar, T., Rodriquez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L. and Boyer, H. B. (1977). Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene *2*, 95-113.

Brandhorst, B. P. (1976). Two-dimensional gel patterns of protein synthesis before and after fertilization of sea urchin eggs. Dev. Biol. *52*, 310-317.

Chen, T. T. and Hodgetts, R. B. (1974). The appearance of dopa decarboxylase activity in imaginal discs of Sarcophaga bullata, undergoing development in vitro. Dev. Biol. *38*, 271-284.

Corces, V., Holmgren, R., Freund, R., Morimoto, R. and Meselson, M. (1980). Four heat shock proteins of Drosophila melanogaster coded within a 12-kilobase region in the chromosome subdivision 67B. Proc. Nat. Acad. Sci. USA *77*, 5390-5393.

Davidson, N., Fyrberg, E. A., Hershey, N. D., Kindle, K., Robinson, R. R., Sodja, A. and Yen, P. (1980). Recombinant DNA studies of DNA sequence organization around actin and tRNA genes of Drosophila melanogaster. In Genetics and Evolution of RNA Polymerase, tRNA, and Ribsomes. S. Osawa, H. Ozeki, H. Uchida and T. Yura, eds. (New York: Elsevier/North-Holland Biomedical Press), pp. 279-295.

Efstratiadis, A., Maniatis, T., Kafatos, F. C., Jeffrey, A. and Vournakis, J. N. (1975). Full length and discrete partial reverse transcripts of globin and chorion mRNAs. Cell *4*, 367-378.

Ephrussi, B. and Beadle, G. W. (1936). A technique of transplantation for Drosophila. Am. Nat. *70*, 218-225.

Fragoulis, E. G. and Sekeris, C. E. (1975). Translation of mRNA for 3,4-dihydroxyphenylalanine decarboxylase isolated from epidermis tissue of Calliphora vicina R. D. in a heterologous system. Eur. J. Biochem. *51*, 305-316.

Fristrom, J. W., Hill, R. J. and Watt, F. (1978). The procuticle of Drosophila: heterogeneity of urea-soluble proteins. Biochemistry *19*, 3917-3924.

Fyrberg, E. A., Kindle, K. L, Davidson, N. and Sodja, A. (1980). The actin genes of Drosophila: a dispersed multigene family. Cell *19*, 365-378.

Gall, J. and Pardue, M. (1971). Nucleic acid hybridization in cytological preparations. Meth. Enzymol. *21*, 470-480.

Griffin-Shea, R., Thireos, G., Kafatos, F. C., Petri, W. H. and Villa-Komaroff, L. (1980). Chorion cDNA clones of D. melanogaster and their use in studies of sequence homology and chromosomal location of chorion genes. Cell *19*, 915-922.

Hepburn, H. R., ed. (1976). The Insect Integument. (New York: Elsevier Scientific Publishing Co).

Hershey, N. D. and Davidson, N. (1980). Two Drosophila melanogaster $tRNA^{Gly}$ genes are contained in a direct duplication at chromosomal locus 56F. Nucl. Acids. Res. *21*, 4899-4910.

Hirsh, J. and Davidson, N. (1981). The isolation and characterization of the dopa decarboxylase gene of Drosophila melanogaster. Mol. Cell Biol., in press.

Jackle, H. (1979). Visualization of proteins after isoelectric focusing during two-dimensional gel electrophoresis. Anal. Biochem. *98*, 81-84.

Kaback, D. B., Angerer, L. M. and Davidson, N. (1979). Improved methods for the formation and stabilization of R-loops. Nucl. Acids Res. *6*, 2499-2517.

Kaback, D. B., Rosbash, M. and Davidson, N. (1981). Determination of cellular RNA concentrations by electron microscopy of R-loop containing DNA. Proc. Nat. Acad. Sci. USA, in press.

Karlson, P. and Sekeris, C. E. (1962). Zum tyrosinstoff Wechsel ter Insekten IX. Kontrolle des Tyrosinstoffwechsels durch Ecdyson. Biochem. Biophys. Acta *63*, 489-495.

Kessler, S. W. (1976). Cell membrane antigen isolation with the staphylococcal protein A-antibody adsorbant. J. Immunol. *117*, 1482-1490.

Keyser, J. W. (1964). Staining of serum glycoproteins after electrophoretic separation in acrylamide gels. Anal. Biochem. *9*, 249-252.

Kraminsky, G. P., Clark, W. C., Estelle, M. A., Gietz, R. D., Sage, B. A., O'Connor, J. D. and Hodgetts, R. B. (1980). Induction of translatable mRNA for dopa decarboxylase in Drosophila: an early response to ecdysterone. Proc. Nat. Acad. Sci. USA *77*, 4175-4179.

Maniatis, T., Jeffrey, A. and Kleid, D. G. (1975). Nucleotide sequence of the rightward operator of phage lambda. Proc. Nat. Acad. Sci. USA *72*, 1184-1188.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. Cell *15*, 687-701.

Mullins, J. I., Casey, J. W., Nicolson, M. O., Burck, K. B. and Davidson, N. (1981). Sequence arrangement and biological activity of cloned FeLV proviruses from a virus productive human cell line. J. Virol. *38*, in press.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250*, 4007-4021.

Rao, R. N. and Rogers, S. G. (1979). Plasmid pKC7: a vector containing ten restriction endonuclease sites suitable for cloning DNA segments. Gene *7*, 79-82.

Ricciardi, R. P., Miller, J. S. and Roberts, B. E. (1979). Purification and mapping of specific mRNAs by hybridization-selection and cell-free translation. Proc. Nat. Acad. Sci. USA *76*, 4927-4931.

Rozek, C. E. and Timberlake, W. E. (1979). Restriction endonuclease mapping by crossed contact hybridization: the ribosomal RNA genes of Achlya ambisexualis. Nucl. Acids Res. *7*, 1567-1578.

Schachat, F. H. and Hogness, D. S. (1973). Repetitive sequences in isolated Thomas circles. Cold Spring Harbor Symp. Quant. Biol. *38*, 371-381.

Spradling, A. C., Digan, M. E., Mahowald, A. P., Scott, M. and Craig, E. A. (1980). Two clusters of genes for major chorion proteins of Drosophila melanogaster. Cell *19*, 905-914.

Taylor, J. M., Illmensee, R. and Summers, J. (1976). Efficient transcription of RNA into DNA by avian sarcoma virus polymerase. Biochem. Biophys. Acta *442*, 324-330.

Wensink, P. C., Finnegan, D. J., Donelson, J. E. and Hogness, D. S. (1974). A system for mapping DNA sequences in the chromosomes of Drosophila melanogaster. Cell *3*, 315-325.

Yen, P. H. and Davidson, N. (1980). The gross anatomy of a tRNA gene cluster at region 42A of the D. melanogaster chromosome. Cell *22*, 137-148.

Yen, P., Hershey, N. D., Robinson, R. and Davidson, N. (1979). Sequence organization of Drosophila tRNA genes. In ICN-UCLA Symposium on Eucaryotic Gene Regulation, R. Axel and T. Maniatis, eds. (New York: Academic Press), pp. 133-141.

# CHAPTER 2

# Cuticle Protein Genes of Drosophila: Structure, Organization, and Evolution of Four Clustered Genes

**Michael Snyder, Michael Hunkapiller, David Yuen, Donald Silvert[+‡], James Fristrom[+], Norman Davidson[*]**

Division of Biology and Department of Chemistry[*], California Institute of Technology, Pasadena, California 91125 and

Department of Genetics[+], University of California, Berkeley, California 94720

Running title: Drosophila cuticle genes

[‡] Present address: Division of Biology, California Institute of Technology, Pasadena, California 91125

Published in Cell **29**, 1027-1040 (1982).

## Summary

A major portion of a 9 kb region of the Drosophila genome containing genes for several cuticle proteins (Snyder et al., 1981) has been sequenced. Five cuticle protein gene-like sequences have been identified and mapped. Amino acid sequences of four of the five major urea-soluble third instar cuticle proteins have been determined. These four sequences are identical with those predicted from the sequences of four of the five genes. Two cuticle genes are transcribed in one direction while two are transcribed in the opposite direction. The fifth cuticle-like gene is judged to be a pseudogene by several criteria; several features of its structure and the absence of detectable transcripts suggest that it is non-functional. Sequence comparisons indicate that it arose by an unequal crossing over event involving two closely related and adjacent cuticle genes. The structures of the four cuticle genes have several interesting features. Each contains a signal peptide coding sequence which is interrupted by a short intervening sequence (about 60 bp) at a conserved site. Conserved sequences occur in the 5' mRNA untranslated region, in the adjacent 35 bp of upstream flanking sequence, and at -200 bp from the mRNA start position in each of the cuticle genes. We discuss the structure, organization and evolution of these cuticle genes as a model small multigene family.

## Introduction

Our knowledge of the mechanisms of gene expression has benefitted by understanding the structure of genes and how they are organized. In bacteria, genes that are coordinately expressed are often organized into operons (see Miller and Reznikoff, 1980), thus allowing coordinate regulation of expression of many genes by relatively few control points. In higher organisms structural genes that are expressed at the same time are often but not always clustered. (For example in Drosophila: chorion genes, Spradling et al., 1980, Griffin-Shea et al., 1980, and Spradling, 1981; two yolk protein genes, Barnett et al., 1980, Riddell et al., 1981; several heat-shock genes, Corces et al., 1980; 68C salivary glue genes, Meyerowitz and Hogness, 1982; histone genes, Goldberg and Hogness, unpublished; cuticle genes, Snyder et al., 1981.) However, in eucaryotes genes are typically separated by large amounts of spacer DNA and there is no evidence of polycistronic messages. Thus, the mechanisms by which eukaryotic genes are coordinately expressed are unknown.

We are studying the larval cuticle protein genes of Drosophila as an example of a set of coordinately expressed genes. Five major urea-soluble cuticle proteins plus a number of minor species are synthesized and secreted by the epidermal cells of late third instar larvae (Fristrom et al., 1978). These proteins and chitin are major components of the cuticle that surrounds and protects the animal through its third instar and pupal stages. Using recombinant DNA techniques, genes for three and probably four of the major third instar cuticle proteins were found to be clustered in a small (7.9 kb) segment of the Drosophila genome (Snyder et al., 1981), at region 44D on the second chromosome, in agreement with previous genetic mapping data (Fristrom et al., 1978; Chihara, Kimbrell and Fristrom, unpublished). The genes encoded in this cluster are abundantly expressed in the integument of third instar larvae and are not abundantly expressed at other developmental stages (Snyder et al., 1981; Snyder and Davidson, unpublished). Genes for the fifth major cuticle protein and several of the less abundant species lie on the third chromosome (Chihara, Hoffman, Kimbrell, and Fristrom, unpublished).

In order to examine further the structural basis for the mechanisms by which the cuticle genes encoded at 44D are expressed coordinately, we have studied their structure and organization by DNA and protein sequencing. These results which include the novel finding of a Drosophila pseudogene are reported below.

## Results and Discussion

### Organization of the Cuticle Gene Cluster

The organization within the cloned region of genes encoding late third instar larval messages as determined in our previous studies (Snyder et al., 1981) is shown in Figure 1. It was shown that genes I and II encode two of the five major cuticle proteins CP1 and CP2, respectively. One of genes III and IV encodes cuticle protein CP3 and the other probably encodes CP4. The last gene, gene V, encodes a nonabundant third instar larval RNA; its identity is unknown.

In order to learn more about the structure of the cuticle genes and their organization within the cluster, we have sequenced the region encoding them and flanking them. DNA fragments from three lambda clones were subcloned (Figure 1) and then sequenced (Maxam and Gilbert, 1980). The strategy used is presented in Experimental Procedures (Figure 7). A total of 8019 bp of DNA from an 8.8 kb region was sequenced; the remaining 0.8 kb segment lies in the spacer region between genes I and II. Greater than 90% of the DNA sequence was determined from both DNA strands in the coding and nearby (300 bp) flanking regions of all four genes and in the gene I/II and gene II/III spacers. Forty percent of the spacer sequences around gene IV were determined from both DNA strands. The entire sequence is presented in the appendix.

Coding regions for late third instar larval messages were localized on the DNA sequence by electron microscopic R-looping and by using labelled cDNA to probe cloned DNA restriction fragments (Snyder et al., 1981). More precise localization of the mature protein coding sequence was possible using the protein sequence data presented below. The results reveal that of the four coordinately expressed genes, I through IV, I and II are transcribed in one direction, whereas III and IV are transcribed in the opposite direction (Figure 1). The distance between these genes from the proposed mRNA start and end positions (see below) is 2.9 kb for genes I and II, 870 bp for genes II and III, and 1600 bp for genes III and IV.

### The Protein Coding Sequence

In order to confirm and extend previous identifications of genes I, II, III and IV (Snyder et al., 1981), the amino terminal sequences of cuticle proteins CP1, CP2, CP3 and CP4 were determined. Between 56 to 72 amino acid residues were sequenced from each protein; this constitutes 51 to 75% of the total in vivo protein sequence. As shown in Figure 2, the DNA sequences of genes I, II, III and IV match perfectly the amino acid sequences of CP1, CP2, CP3 and CP4, respectively. Since there are no other copies of these genes outside this cluster (Snyder et al., 1981), genes I through IV must encode these proteins.

The proteins were isolated from a Drosophila melanogaster Oregon-R wild type strain, however the sequenced DNA was from Canton S-derived clones. Comparison of the Oregon-R protein sequence with that deduced from the Canton S DNA sequence showed no amino acid polymorphism between these two strains in the 250 amino acid residues examined.

Further comparison of the protein and DNA sequences reveals that each of the four cuticle genes encodes a signal peptide of fifteen amino acid residues (Figure 3). This signal peptide is comprised of hydrophobic amino acid residues except for one lysine residue two positions downstream from the initiator methionine. Similar structures are found for most secreted proteins (see Kreil, 1981). In all four genes a small intron occurs between the sequences encoding the third and fourth amino acid of the signal peptide (shown below).

Analysis of each of the protein sequences for its hydrophobic and hydrophilic regions (Kyte and Doolittle, 1982) shows that, while most of the protein is acidic and somewhat hydrophilic there are two hydrophobic regions: one in the signal peptide sequence and the other at the carboxy terminus of the protein (approximate nucleotide positions 268-301 in Figure 2). This latter region directly follows a proline-rich sequence (5 out of 7 residues are proline).

**Cuticle Genes Contain an Intron at a Conserved Position.**

We surmised that there was an intron in the 5' end of each of these genes for the following reasons: First, translation of proteins usually begins at the first AUG of the mRNA (Kozak, 1978), which is the one indicated in Figure 2. (The mRNA start position is discussed below.) For genes I through III the first AUG is out of frame with the in vivo protein coding sequence; and for all four cuticle genes this AUG is followed by in-frame termination codons 7 to 17 codons downstream. Second, for genes III and IV the predicted protein sequence immediately upstream from the mature protein terminus shows no AUGs but does show termination codons. Finally, inspection of the DNA sequence for each of the four genes reveals a consensus RNA splicing donor site 3 codons after the first AUG in the mRNA; 50 to 60 bp away lies a consensus splicing acceptor site, 11 codons upstream from the mature protein coding sequence (Figures 2, 3b). We therefore predict the presence of a short intron in a conserved position for each of genes I, II, III and IV (Figure 3b); RNA splicing using the consensus splicing sequence would bring an initiation codon and part of the signal peptide coding sequence into frame with the remainder of the coding sequence for the signal peptide and that of the mature protein.

The occurrence of introns was directly demonstrated and the mRNA start positions were identified for genes I and III by comparing the genomic sequences with the sequences of cDNA copies of the respective mRNAs. This was accomplished by preparing labeled DNA primers which are homologous to the mRNA encoding the amino terminus of the in vivo protein and part of the signal peptide (Figure 3c). DNA primers were chosen so as to minimize homology to RNAs from other cuticle genes (see below). Each primer was hybridized to total late third instar larval RNA and cDNA was synthesized. The length of such a cDNA for gene I is shown in Figure 3a. Two predominant species are observed, which differ in length by one nucleotide. The ends of these cDNAs (* in Figure 3c) map to a conserved sequence (discussed below) 24-25 bp downstream from the TATA or Goldberg-Hogness box, after allowing for the length of the proposed intron. A minor cDNA species (1% relative to the major species) is twice as long as the major species, and is presumably the result of double stranded cDNA synthesis by reverse transcriptase. When a primer for gene III was used, an identical result was observed (Figure 3c). For both of these genes the first nucleotide of the mRNA is probably the A residue rather than T (see Figure 3c), since most eucaryotic mRNAs contain an adenine residue after their 5' cap (Corden et al., 1980).

We directly demonstrated the presence of introns by sequencing the cDNAs formed for genes I and II. The result is shown in Figure 3c; the intron sequences predicted are not present in the mature mRNA. Because genes II and IV have a high degree of homology with genes I and III, respectively (discussed below), we expect both the mRNA start sites and the intron positions to be located at the analogous positions. It is unlikely that any other introns lie in the protein coding sequence of these genes for several reasons. First, between 51 and 75% of the proteins encoded by genes I through IV has been sequenced and the protein sequences are colinear with the DNA sequences. In the remaining DNA sequence, where the protein has not been sequenced, no GTPurine triplet, which is found at all splicing donor sites thus far studied (Lerner et al., 1980; Figure 3b), is present. (Genes III and IV do not even contain a GT within these sequences.) Second, the amino acid compositions of CP1, CP2, CP3 and CP4 (Fristrom et al., 1978) agree remarkably well with those predicted from the DNA sequences of genes I, II, III and IV. (From our work, however, it appears that the His and Arg contents previously reported for CP1 and CP2 are reversed.) In addition, the proteins encoded by these four genes are approximately the size expected from the DNA sequences (Fristrom et al., 1978; Snyder et al., 1981). These latter observations further indicate that large segments of the proteins are not cleaved post-translationally. Finally, the protein coding sequences predicted from the DNA sequences for these genes are remarkably homologous (see below); more divergent regions might be expected if intervening sequences were present. Our data allow no conclusions about the presence of introns in the 3' untranslated region.

## The Cuticle Genes Within the Cluster are Homologous and Comprise a Small Multigene Family

### A. Protein Coding Sequence Homology

As seen in Figure 2 and Table 1, genes I and II are very homologous (91%) in the DNA sequences encoding the mature proteins. At the protein level CP1 and CP2 differ only in six residues close to the amino terminus. Similarly, genes III and IV are closely related in their in vivo protein coding regions (85% in DNA sequence). These similarities are consistent with the observation that gene I cross-hybridizes with gene II, and gene III cross-hybridizes with gene IV under moderately stringent conditions (Snyder et al., 1981). In addition, between genes I or II with III or IV, there is DNA sequence homology ranging from 59 to 62% depending upon the comparison made (Table 1). In general, the amino terminal sequences of these proteins are less well conserved than the remainder of the protein sequences.

Homology comparisons have revealed several other interesting features. First an imperfect gene duplication of 33 bp exists near the amino termini for genes I and II (bp positions 22 to 54 and 55 to 87 in Figure 2); only part of one copy is found in genes III and IV (positions 67 to 87). The duplicated segment in genes I and II contains 7 and 8 conserved amino acid residues, respectively. Second, a 12 bp insertion occurs at or close to the amino terminus of gene I relative to gene II. Finally, comparison of the in vivo coding sequence of CP3 and CP4 reveals that in the 12 amino acid positions where these proteins differ, 5 residues in CP4 are identical to those of genes I and II. This latter observation is discussed below.

B. 5' Untranslated, Intervening Sequence and Signal Peptide Coding Region Homology

The signal peptide protein sequences are less well conserved than the mature protein sequences, except for the comparison of genes III and IV (Table 1); and little homology is found within the intervening sequences except at the splice junction borders (Table 1; Figure 3b). However, a high degree of sequence conservation occurs in the entire mRNA 5' untranslated region (Figure 3c). In particular, there is an identical start sequence ATCAGTC in each of these genes (Figure 3c). Related sequences are found at similar positions for other Drosophila genes including yolk protein 1 (GCCAGTT, Hovemann et al., 1981), heat shock proteins (22k-CTCAGTT, 23k-GTCAGTT, 26k-CACAGAT, 27k-CACAGTC, 83k-TCGAGTC, see Ingolia and Craig, 1981; Holmgren et al., 1981), 68C gene products (II-ATCAGTT, III and IV-ATCTGGT; Garfinkel, Pruitt and Meyerowitz, in preparation) and actin (79B-ATCACTC; Sánchez et al., submitted). Furthermore, imperfect repeats of the ATCAGTC sequence are present immediately upstream of the mRNA start site for genes I and II. The cDNA synthesis experiments indicate that only the ATCAGTC sequence, which lies 24-25 nucleotides downstream from the TATA box, is used. These results suggest that in order to function as a start sequence, the consensus sequence must lie at an appropriate distance from the TATA box. Previous studies also indicate that initiation of transcription occurs at a preferred position relative to the TATA box (McKnight et al., 1981; Grosschedl and Birnstiel, 1980).

### C. Upstream Sequences

In the first 35 bp of upstream flanking sequences there is extensive DNA sequence homology (-1 to -35, Figure 4a) between the cuticle genes. This region includes the consensus TATA or Goldberg-Hogness box 24-25 residues upstream from the mRNA start position. Upstream from the TATA box the DNA sequence is AT-rich and several short homologous sequences in this region are depicted in Figure 4a. In the region from -60 to -85, genes I and II contain several sequences similar to the CAAT sequence observed in other eukaryotic genes (Benoist et al., 1980); we do not discern similar sequences in genes III and IV. However, these latter genes have identical sequences, TGCATCA, starting at position -76; related sequences are found in the same region for genes I and II, as indicated by boxes in Figure 4a. At the -200 position another homologous sequence is found of length 14 bp in genes I, II and III and 9 bp in gene IV.

### D. 3' Untranslated Region

A consensus poly(A) addition sequence, AATAAA (Proudfoot and Brownlee, 1976), for genes III and IV lies 110 bp away from the translation termination codon (Figure 4b). Based on precedents from other genes poly(A) addition probably occurs approximately 20 residues after this sequence. For genes I and II there are similar, but not identical, sequences in the same region. Therefore we predict the 3' untranslated regions to be 110-140 residues long, which makes the total mRNA sizes consistent with previous measurements (Snyder et al., 1981). These 3' flanking regions contain several short homologous segments, shown in Figure 4b, but overall they appear less conserved than either the mature protein coding regions or the 5' untranslated regions.

## A Putative Pseudogene Lies Within This Cluster

500-600 bp downstream from gene II lies another cuticle gene-like sequence which we denote as ψ gene I. The single long open reading frame has a sequence that is quite closely related to that of genes I and II (Figure 5, Table 1). The 5' protein coding region, intervening sequence and upstream flanking sequence of ψ gene I are all much more homologous to gene I than to gene II. (Compare Figure 5a panel i with panel ii.) In contrast, a short region at the 3' end shows more extensive homology to gene II then gene I (Figure 5a and 5b). Further homology comparisons indicate that 170–220 bp upstream from ψ gene I and gene I there is a 500 bp DNA insertion in gene I relative to ψ gene I. Alternatively, this feature may have resulted from a 500 bp deletion in ψ gene I. Similarly, at the 3' end of ψ gene I, there is a 180 bp deletion in ψ gene I relative to gene II (Figure 5a).

We believe that ψ gene I is a pseudogene for several reasons. First, a 35 bp deletion has eliminated the region encoding the TATA box (Figure 5b). The TATA box region has been demonstrated to be necessary for efficient transcription of several genes (Corden et al., 1980; Hu and Manley, 1981; Grosveld et al., 1981; Dierks et al., 1981; Wasylyk et al., 1980; Wasylyk and Chambon, 1981) and for selecting the proper mRNA start position in other genes (Rio et al., 1980; Mathis and Chambon, 1981; Myers et al., 1981; Gluzman et al., 1980; Benoist and Chambon, 1981; Ghosh et al., 1981; Grosveld et al., 1981; McKnight et al., 1981). Second, although ψ gene I has a conserved translation initiation codon and splicing donor sequence, the splicing acceptor sequence at the position analogous to genes I-II is both out of frame and mutated such that it no longer contains an AG sequence. A PyrimidineAG sequence is found at all eukaryotic splice acceptor boundaries (Figure 3b). Downstream there are no other in-frame PyrimidineAG sequences; however, a splicing event would be necessary, since no translation initiation codons are in frame with the protein coding sequence of ψ gene I. Finally, there are several substitutions, insertions and deletions (Figure 5b) some of which would have striking effects on the proteins produced: A TGA codon near the COOH end of the protein-coding sequence would cause termination 19 amino acid residues earlier than in genes I and II. The several codons for charged amino acids in the signal peptide coding region would be expected to prevent secretion of the protein. Furthermore, ψ gene I exhibits the same proportion of replacement site mutations as silent mutations, a result expected for a nonfunctional gene (Figure 5b) (Efstratiadis et al., 1980).

Additionally, no abundant transcripts for ψ gene I are found in two embryonic, second instar larval, late third instar larval, four pupal and one adult stages, or in imaginal discs (Snyder et al., 1981; Snyder and Davidson, unpublished). Since the predicted protein sequence is cuticle protein-like, we have included in our examination all stages where cuticle deposition occurs (Chihara et al., 1982) except for the first instar larval stage (which makes the same cuticle proteins as the second instar stage). Our experiments do not exclude the possibility that ψ gene I is transcribed at a very low level; less than $10^{-5}$ of total poly(A) RNA in late third instar larvae (Snyder et al., 1981).

**Spacer Regions**

In general, spacer segments between eukaryotic genes are AT-rich. The spacer DNA between the four cuticle genes exhibits this same pattern. We further note the absence of any open reading frame greater than 250 bp.

## Further Discussion

### Structure, Expression and Evolution of the Cuticle Gene Cluster.

Four of the cuticle protein genes that are expressed coordinately in third instar larvae are clustered within a 7.9 kb region of the Drosophila genome. The clustering of genes may contribute to the mechanism of their coordinate expression; such a mechanism may involve the activation of domains of chromatin for transcription. Alternatively, as discussed below, genes I through IV probably arose through gene duplication events; they may not yet have had time to disperse into other parts of the genome as known for other multigene families (for example, actin, Fyrberg et al., 1980; Tobin et al., 1980; α and β tubulins, Sánchez et al., 1980; and larval serum protein 1, Smith et al., 1981).

Additionally, we have noted several conserved sequences in the 5' mRNA untranslated region and upstream flanking sequences as well as at -200 bp (one nucleosome) from the mRNA start site. These sequences may play an important role in controlling the expression of these genes. Sequence conservation in the 5' mRNA ends and flanking regions of Drosophila heat shock genes has also been found (Ingolia and Craig, 1981; Karch et al., 1981; Holmgren et al., 1981).

Each cuticle gene contains an intron 56 to 64 bp long at a conserved position within its signal peptide coding region. Short introns (<100 bp) are quite common in Drosophila genes, examples being actin 88F (Sánchez et al., submitted), alcohol dehydrogenase (which has two) (Benyajati et al., 1980), 83K heat shock protein gene (Holmgren et al., 1981), yolk protein gene I (Hovemann et al., 1981), 3 larval serum protein 1 genes (McClelland et al., 1981), and 3 glue protein genes encoded at 68 C (M. D. Garfinkel, R. E. Pruitt, and E. M. Meyerowitz, in preparation). The small size and number of introns in Drosophila genes may be related to the manner in which Drosophila economizes its DNA relative to other organisms (discussed below). From the limited data available, introns in Drosophila structural genes occur more frequently near the 5' end than in the rest of the gene (see references for sequenced genes listed in Figure 3b plus dopa decarboxylase, Hirsh and Davidson, 1981 (unsequenced)).

Sequence comparisons of the larval cuticle genes suggest a simple scheme as to how this cluster may have arisen (Figure 6). Duplication events produced the cuticle genes. Genes III and IV or their precursors must have inverted their orientation relative to genes I and II (or precursors). An additional event that occurred is the duplication of a 33 bp gene segment within genes I and II or their precursor, relative to genes III and IV.

Genes I and II have the same orientation. ψ gene I was probably formed by unequal crossing between genes I and II. The detailed sequence comparisons suggest that ψ gene I is derived from gene I in the region upstream of nucleotide 468 (Figure 5b), which contains the 5' flanking sequence, intron and most of the protein coding sequence; ψ gene I is derived from gene II in the region downstream from nucleotide 499, which contains the putative 3' untranslated sequence and flanking sequences. It subsequently evolved into a pseudogene. There is a relatively low degree of sequence divergence of ψ gene I from its proposed parental sequences of gene I and II even in the flanking regions and intron; this observation indicates that the formation of ψ gene I was the most recent event in the evolution of the cluster.

Genes III and IV are inverted relative to I and II. For Drosophila, divergent orientation of homologous genes is rather common, having been found for histone genes (Goldberg and Hogness, unpublished), two yolk protein genes (Barnett et al., 1980; Riddell et al., 1981), 68C genes (Meyerowitz and Hogness, 1982; M. D. Garfinkel, R. E. Pruitt and E. M. Meyerowitz, in preparation), the 67B heat shock protein genes (Corces et al., 1980) and the 70k heat shock protein genes at both 87A and 87C (see Brown and Ish-Horowicz, 1981). For gene clusters in organisms with larger genome sizes, arrangement of genes in the same orientation seems to be more frequent (α globin genes in human: Lauer et al., 1980; β globin genes in mouse, rabbit, goat, chicken, human: Leder et al., 1980; Jahn et al., 1980; Lacy et al., 1979; Cleary et al., 1980; Bernards et al., 1979; Dolan et al., 1981; Fritsch et al., 1980; ovalbumin, X and Y genes: Royal et al., 1979; sea urchin histone genes: see Hentschel and Birnstiel, 1981; several sea urchin actin genes: Scheller et al., 1981; mouse major histocompatibility genes: Steinmetz et al., 1982; mouse α fetoprotein and serum albumin genes: Ingram et al., 1981).

The inverted orientation of closely spaced homologous genes may place several useful constraints on DNA evolution. This arrangement prevents gene duplication by unequal crossing over thus inhibiting gene correction mechanisms and leaving the gene sequences free to diverge. We note that genes III and IV are divergent from I and II in sequence. In addition, Drosophila melanogaster has the smallest genome known for higher organisms (Figure 1 of Britten and Davidson, 1971). We propose that the maintenance of a small genome size constrains the proliferation of non-essential sequences, including pseudogenes and long introns. Because gene inversion prevents gene duplication and thereby provides a mechanism for preventing pseudogene formation, it may be useful in maintaining the small size of the Drosophila genome. In fact, pseudogenes occur infrequently in Drosophila (ψ gene I is the first known example for a structural gene). In contrast, for mammals there are many examples of pseudogenes for structural genes including α globin genes (Proudfoot and Maniatis, 1980; Nishioka et al., 1980; Vanin et al., 1980; Lauer et al., 1980), β globin genes (Cleary et al., 1980; Fritsch et al., 1980; Hardison et al., 1979; Jahn et al., 1980; Lacy and Maniatis, 1980), heavy chain variable region genes (Huang et al., 1981), and major histocompatibility genes (Steinmetz et al., 1981). Finally, the divergent orientation of homologous genes may decrease the frequency of excision by intrachromosomal recombination.

Detailed comparisons of the cuticle genes show that even though genes III and IV are closely related, IV is more similar to I and to II than is III. Since III and IV arose from a more recent duplication than the duplication that led to the I/II and III/IV pairs, this argues that selective pressure is operating to maintain the sequence of CP4. Consistent with this hypothesis is the observation that no wild type Drosophila melanogaster strains have been found that contain electrophoretic variants of CP4. In contrast, strains making electrophoretic variant proteins for each of the other major cuticle proteins have been found (Fristrom et al., 1978).

**Cuticle Proteins and mRNAs**

In the protein sequencing studies we found no evidence of glycosylation, which is in agreement with previous results by other methods (Snyder et al., 1981; Silvert and Fristrom, submitted), nor of N-terminal acetylation. Furthermore, two dimensional gels reveal no protein phosphorylation or acetylation (Snyder et al., 1981). Hence, other than signal peptide processing, no initial posttranslational modification of these cuticle proteins is evident. However, for many insects evidence indicates that cuticle proteins, once secreted into the cuticle, become cross-linked at pupariation during the tanning of the cuticle (sclerotization) (reviewed in Hepburn, 1976).

We previously noted that more than two proteins are produced from in vitro translation of RNA selected by genes I and II (Snyder et al., 1981). We have now resolved four in vitro translation products of these two genes (data not shown). The isoelectric focusing patterns and methionine labeling patterns of these proteins are consistent with the interpretation that in vitro translation begins at both of the closely spaced AUGs (-48 and -34) for each of genes I and II (Figure 2). Thus, the second AUG is not excluded from use for cuticle gene messages in a rabbit reticulocyte translation system. In addition, we previously noted the occurrence of the two translation products, t2V and t3V, from mRNA of the 2/3 cuticle variant Drosophila strain (Snyder et al., 1981); these may also be due to initiation of translation from two closely spaced AUGs.

## Experimental Procedures

### Lambda and Plasmid Clones

The isolation of lambda clones was previously described (Snyder et al., 1981). The DNA fragments depicted in Figures 1 and 7 were subcloned according to Snyder et al. (1981). All fragments were ligated into appropriate sites in the vector pBR322 (Bolivar et al., 1977) except that Bgl II-cut DNA fragments were ligated into the plasmid pKC7 (Rao and Rogers, 1979). The plasmid pCPII/III-2 contains both the fragment indicated in Figure 1 plus an additional 0.6 kb Bgl II fragment from the vector Charon 4. All plasmid subclones were maintained in E. coli strain HB101 (Boyer and Roulland-Dussoix, 1969).

### Nucleic Acid Preparations

DNAs and undergraded cellular RNA from late third instar Drosophila larvae were prepared as described (Snyder et al., 1981).

### DNA Sequencing

The sequencing strategy used is depicted in Figure 7. DNA fragments were labeled either at 5' termini by T4 polynucleotide kinase plus $\gamma^{32}$P ATP or by filling in 3' termini with $\alpha^{32}$P NTPs using E. coli polymerase I large fragment, according to standard protocols (Maxam and Gilbert, 1980). The DNA fragments were sequenced according to the procedures of Maxam and Gilbert (1980). Six reactions were used: G, G>A, G+A, A>C, T+C, C, except that in 80% of the pCPIV-8 sequencing, the G>A reaction was omitted. Initially, cleavage products were analyzed on 40 cm gels. Later, we used 80 cm gels and shorter reaction times to read the ladders further (Smith and Calvo, 1981). For the longer gels the Maxam and Gilbert protocol (1980) was modified as follows: G and G>A, 0.5 μl dimethylsulfate 3-4 min at 20°C; G+A 80 min at 20°C; A>C 6 min at 90°C; T+C and C, 5 min at 20°C. The products were analyzed on 25%, 8% and 5% acrylamide gels; in several cases over 500 nucleotides could be read from a single end labeled fragment. The sequence was checked by restriction endonuclease digestions with Taq I, Hpa II, Sau 3A, Hinf I, Ava I, Ava II, Pst I, Xba I, Sph I, Hpa I plus all enzymes shown in Figure 1 and in some regions with Rsa I, Sau 96A, Fnu 4HI, Hae III and Bst NI. In spacer regions where only one strand was sequenced the ladder appeared reliable and the frequency of errors is estimated to be less than 1%. All Eco RII sites appear methylated and do not cleave at C (Maxam and Gilbert, 1980). In addition, the A positions at Mbo I sites often appeared fainter than at unmethylated positions, particularly when the modified reactions were used. These deficiencies were checked by both sequencing the complementary DNA strand and/or by digestion with either Bst NI or Sau 3A. In addition, we note that two Ava II sites predicted by the DNA sequence that partially overlap Eco RII sites ($\mathrm{CC}^{A}_{T}\mathrm{GG}^{A}_{T}\mathrm{CC}$) do not cleave with Ava II under standard conditions. This is presumably due to methylation at these sites. Finally, at positions 2938 and 2945 in the Appendix, two closely spaced Hpa II sites are predicted by the DNA sequence; however, cleavage appears to only occur at the 2945 position. Several discrepancies were found when sequencing both DNA strands, particularly at the 5' end of gene I where several palindromic sequences reside. The 5' ends of genes II and IV gave similar problems. The discrepancies were resolved by sequencing these regions from three or four different positions; the most reliable sequence is presented. Finally, we point out that sequencing was performed across all restriction endonuclease sites except for a Bgl II site and an Eco RI site in the III/IV spacer.

### mRNA Sequencing by cDNA Synthesis and Sequencing

a) DNA primer preparation. Labeled DNA primers were prepared as follows: Gene I: 10 μg of plasmid pCPI-11 was cleaved with Hae III and the mixture of fragments was labeled at their 5' ends with $^{32}$P γ ATP plus T4 polynucleotide kinase. The 28 bp primer fragment was isolated on a 20% acrylamide gel according to Maxam and Gilbert (1980). It contains 5' label on both the coding and noncoding strand. This fragment is not homologous with gene II (see Figures 5c and 3). Gene III: a 464 bp Hpa II fragment was isolated on a 5% acrylamide gel from 10 μg of pCPIII-9. The fragment was labeled with T4 polynucleotide kinase plus $^{32}$P γ ATP, recleaved with Hae III and the 111 bp DNA primer was isolated. This fragment is labeled at only one end and is 83% homologous with gene II.

b) Hybridization. The labeled gene I and III primer DNA fragments (1-5 $10^{6}$ DPM) were hybridized to 500 and 750 μg total RNA from late third instar larvae, respectively, in 150 μl of 70% formamide, 0.4 M NaCl, 0.1 M PIPES, pH 6.4, 1 mM EDTA at 52°C for 3 to 3.5 hr. Prior to incubation, samples were heated to 70°C for 2 min to denature the DNA primer fragments. After hybridization the solution was diluted with 2 ml of 0.5 M NaCl, 0.01 M Tris-HCl, pH 7.4, 1 mM EDTA, .1% SDS (B. buffer) and passed three times through 0.1 gm oligo(dT) cellulose (Collaborative Research). The column was washed seven times with 1.5 ml B. buffer, then five times with 1.5 ml B. buffer without SDS. The poly$(A)^+$ RNA with bound DNA primer was then eluted with 4 aliquots of 0.5 ml double distilled $H_2O$. The RNA was recovered by ethanol precipitation, and washed with 95% ethanol. The oligo(dT) column step was included to collect the poly$(A)^+$ RNA and remove much of the unhybridized primer and any other contaminating DNA fragments. 80-90% of the labeled material was in the unbound fraction in each case.

c) cDNA synthesis and sequencing. A cDNA was synthesized according to standard techniques (Ghosh et al., 1980). The RNA pellet was dissolved in 150 µl of cDNA buffer: 60 mM NaCl, 50 mM Tris-HCl, pH 8.5, 6 mM $MgCl_2$, 20 mM DTT, 200 mM each of dATP, dTTP, dCTP and dGTP. 3 µl (33 units) of reverse transcriptase (gift of J. Beard) was added and incubated for 3.0 hr at 38°C. The reaction was stopped by incubation with 100 ng of boiled RNase A for 15 min at 37°C. The DNA was recovered by phenol and chloroform extractions, ethanol precipitated, and washed with 95% ethanol twice.

4% of the sample ($5 \times 10^3$ DPM) was dissolved in denaturing loading solution (Maxam and Gilbert, 1980) and analyzed on 8% acrylamide 80 cm sequencing gels. The remainder of the sample was subjected to the six DNA sequencing reactions and analyzed as outlined above. There was no evidence of cross-hybridization of these DNA primers with homologous RNA from other genes; less than 10% contamination would not be detected. In addition, from the amount of DNA primer that hybridized and produced cDNA, we estimate the abundance of gene I and III messages to be 0.3% of total late larval poly$(A)^+$ RNA, a figure in agreement with previous results (Snyder et al., 1981).

**Protein Purification and Sequencing**

Cuticle proteins were extracted from mass-isolated cuticles of third instar Oregon-R larvae and chromatographed on DEAE-cellulose as previously described (Fristrom et al., 1978). Fractions containing CP3, CP1 and CP2, and CP4 and CP5, were recovered and pooled. The proteins were rebound to DEAE-cellulose and stepped from the column in a minimal volume (ca. 5 ml) of 0.1 M NaCl, 7 M urea, 5 mM Tris, pH 8.6. The concentrated material was dialyzed against 7 M urea. Ampholytes (Biorad) were then added to a final concentration of 2% (for CP1, CP2, and CP3) or 3% (for CP4). Preparative isoelectric focusing was carried out in a flat bed apparatus using polyacrylamide beads as the supporting medium. CP3 was focused one time using a 2% mixture of 20% pH 4-6, 80% pH 3-7 ampholytes and was recovered at an approximate pH of 4.5. CP1 and CP2 were focused together using a 2% mixture of 20% pH 5-7, 80% pH 3-7 ampholytes and were recovered at approximately pH's of 5.9 and 5.7, respectively (Snyder et al., 1981). The separate proteins were then refocused using the same conditions. Chromatographic fractions containing CP4 and CP5 were rechromatographed on DEAE-cellulose and fractions on the leading edge of the eluted peak, enriched for CP4, were recovered. This material was focused using a 3% mixture of 50% pH 4-6; 50% pH 3-5 ampholytes. CP4 was recovered from the basic side of a band that focused at pH 4.75. Densitometric analysis after electrophoresis revealed no detectable contaminants of CP1, CP2, and CP3. The CP4 material contained 97% CP4 and 3% CP5. Purified fractions were dialysed against distilled water and lyophilized in preparation for sequence analysis.

Amino terminal amino acid sequence analysis on samples (5-10 nmoles) of the purified cuticle proteins were performed using automated Edman degradation on a spinning cup sequenator according to Hunkapiller and Hood (1980). Polybrene was used as a carrier for the proteins in the sequenator, and the degradation was performed with Quadrol coupling buffer, double cleavage, and automated phenylthiohydantoin conversion (aqueous trifluoroacetic acid). Amino acid phenylthiohydantoins were analyzed by reverse phase HPLC on a DuPont Zorbax CN column.

## Acknowledgements

We are grateful to Tim Hunkapiller for expert help with the computer analyses and to Dr. Russ Doolittle for the analysis of the hydrophobic regions of each of the cuticle proteins. We also thank Dr. Lee Hood in whose laboratory the protein sequencing was performed and we are indebted to Dr. N. Davis Hershey, R. Davis, M. Garfinkel and E. Meyerowitz for criticizing the manuscript.

This research was supported by National Institutes of Health research grants to N.D., M.H. and J.F. and by an NIH predoctoral training grant to M.S. and D.S.

## References

Barnett, T., Pachl, C., Gergen, J. P. and Wensink, P. C. (1980). The isolation and characterization of Drosophila yolk protein genes. Cell **21**, 729-738.

Benoist, C. and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. Nature **290**, 304-310.

Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980). The ovalbumin gene sequence of putative control regions. Nucl.Acids Res. **8**, 127-142.

Bernards, R., Little, P. F. R., Annison, G., Williamson, R. and Flavell, R. A. (1979). Structure of the $G_{\gamma}$-$A_{\gamma}$-δθβ-globin gene locus. Proc. Natl. Acad. Sci. USA **76**, 4827-4831.

Benyajati, C., Place, A. R., Powers, D. A. and Sofer, W. (1981). Alcohol dehydrogenase gene of Drosophila melanogaster: Relationship of intervening sequences to functional domains in the protein. Proc. Nat. Acad. Sci. USA **78**, 2717-2721.

Bolivar, F., Rodriquez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. B., Crosa, J. H. and Falkow, S. (1977). Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2**, 95-113.

Boyer, H. W. and Roulland-Dussoix, D. (1969). A complementation analysis of the restriction and modification of DNA in Escherichia coli. J. Mol. Biol. **41**, 459-472.

Britten, R. J. and Davidson, E. H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. The Quarterly Review of Biology **46**, 111-138.

Brown, A. J. L. and Ish-Horowitz, D. (1981). Evolution of the 87A and 87C heat-shock loci in Drosophila. Nature **290**, 677-682.

Chihara, C. J., Silvert, D. J. and Fristrom, J. W. (1982). The cuticle proteins of Drosophila melanogaster: Stage specificity. Dev. Biol. **89**, 379-388.

Cleary, M. L., Haynes, J. R., Schon, E. A. and Lingrel, J. B. (1980). Identification by nucleotide sequence analysis of a goat pseudoglobin gene. Nuc. Acids Res. **8**, 4791-4802.

Corces, V., Holmgren, R., Freund, R., Morimoto, R. and Meselson, M. (1980). Four heat shock proteins of Drosophila melanogaster coded within a 12-kilobase region in the chromosome subdivision 67B. Proc. Nat. Acad. Sci. USA **77**, 5390-5393.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, P. and Chambon, P. (1980). Promoter sequences of eukaryotic protein-coding genes. Science **209**, 1406-1414.

Dierks, P., van Ooyen, A., Mantei, N. and Weissman, C. (1981). DNA sequences preceding the rabbit β-globin gene are required for formation in mouse L cells of β-globin RNA with the correct 5' terminus. Proc. Nat. Acad. Sci. USA **78**, 1411-1415.

Dolan, N., Sugarman, B. J., Dodgson, J. B. and Engel, J. D. (1981). Chromosomal arrangement of the chicken β-type globin gene. Cell **24**, 669-677.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980). The structure and evolution of the human β-globin gene family. Cell **21**, 653-668.

Fristrom, J. W., Hill, R. J. and Watt, F. (1978). The procuticle of Drosophila: heterogeneity of urea-soluble proteins. Biochemistry **19**, 3917-3924.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980). Molecular cloning and characterization of the human β-like globin gene cluster. Cell **19**, 959-972.

Fyrberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S. and Davidson, N. (1981). The actin genes of Drosophila: Protein coding regions are highly conserved but intron positions are not. Cell **24**, 107-116.

Fyrberg, E. A., Kindle, K. L., Davidson, N. and Sodja, A. (1980). The actin genes of Drosophila: A dispersed multigene family. Cell **19**, 365-378.

Ghosh, P. K., Lebowitz, P., Frisque, F. J. and Gluzman, Y. (1981). Identification of a promoter component involved in positioning the 5' termini of simian virus 40 early mRNAs. Proc. Nat. Acad. Sci. USA **78**, 100-104.

Ghosh, P. K., Reddy, V. B., Piatak, M., Lebowitz, P. and Weissman, S. M. (1980). Determination of RNA sequences by primer directed synthesis and sequencing of their cDNA transcripts. Methods in Enzymology **65**, 580-595.

Griffin-Shea, R., Thireos, G., Kafatos, F. C., Petri, W. H. and Villa-Komaroff, L. (1980). Chorion cDNA clones of D. melanogaster and their use in studies of sequence homology and chromosomal location of chorion genes. Cell **19**, 915-922.

Gluzman, Y., Sambrook, J. and Frisque, R. (1980b). Expression of early genes of origin-defective mutants of simian virus 40. Proc. Nat. Acad. Sci. USA **77**, 3898-3902.

Grosschedl, R. and Birnstiel, M. L. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proc. Nat. Acad. Sci. USA **77**, 1432-1436.

Grosveld, G. C., Shewmaker, C. K., Jat, P. and Flavell, R. A. (1981). Localization of DNA sequences necessary for transcription of the rabbit β-globin gene in vitro. Cell **25**, 215-226.

Hardison, R. C., Butler, E. T. III, Lacy, E. and Maniatis, T. (1979). The structure and transcription of four linked rabbit β-like globin genes. Cell **18**, 1285-1297.

Hentschel, C. C. and Birnsteil, M. L. (1981). The organization and expression of histone gene families. Cell **25**, 301-313.

Hepburn, H. R., ed. (1976). The Insect Integument. (New York: Elsevier Scientific Publishing Co).

Hirsh, J. and Davidson, N. (1981). Isolation and characterization of the dopa decarboxylase gene of Drosophila melanogaster. Mol. Cell. Biol. **1**, 475-485.

Holmgren, R., Corces, V., Morimoto, R., Blackman, R. and Meselson, M. (1981). Sequence homologies in the 5' regions of four Drosophila heat-shock genes. Proc. Natl. Acad. Sci. USA **78**, 3775-3778.

Hovemann, B., Galler, R., Waldorf, U., Küpper, H. and Bautz, E. F. K. (1981). Vitellogenin in Drosophila melanogaster: Sequence of the yolk protein I gene and its flanking regions. Nucl. Acids Res. **9**, 4721-4734.

Hu, S.-L. and Manley, J. (1981). DNA sequence required for initiation of transcription in vitro from the major late promoter of adenovirus 2. Proc. Nat. Acad. Sci. USA **78**, 820-824.

Huang, H., Crews, S. and Hood, L. (1981). An immunoglobulin $V_H$ pseudogene. J. Mol. Appl. Genet. **1**, 93-101.

Hunkapiller, M. and Hood, L. E. (1980). New protein sequenator with increased sensitivity. Science **207**, 523-525.

Ingolia, T. D. and Craig, E. A. (1981). Primary sequence of the 5' flanking regions of the Drosophila heat shock genes in chromosome subdivision 67B. Nucl. Acids Res. **9**, 1627-1642.

Ingram, R. S., Scott, R. W. and Tilghman, S. M. (1981). α-fetoprotein and albumin genes are in tandem in the mouse genome. Proc. Natl. Acad. Sci. USA **78**, 4694-4698.

Jahn, C. L., Hutchinson, C. A. III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F. and Edgell, N. H. (1980). DNA sequence organization of the β-globin complex in the BALB/ mouse. Cell **21**, 159-168.

Karch, F., Töräk, I. and Tissières, A. (1981). Extensive regions of homology in front of the two hsp 70 heat shock variant genes in Drosophila melanogaster. J. Mol. Biol. **148**, 219-230.

Kozak, M. (1978). How do eukaryotic ribosomes select initiation regions in messenger RNA? Cell **15**, 1109-1123.

Kreil, G. (1981). Transfer of Proteins across membranes. Ann. Rev. Biochem. **50**, 317-348.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., in press.

Lacy, E., Hardison, R. C., Quon, D. and Maniatis, T. (1979). The linkage arrangement of four rabbit β-like globin genes. Cell **18**, 1273-1283.

Lacy, E. and Maniatis, T. (1980). The nucleotide sequence of a rabbit β-globin pseudogene. Cell **21**, 545-553.

Lauer, J., Shen, J.C.-K. and Maniatis, T. (1980). The chromosomal arrangement of human α-like globin genes: Sequence homology and α-globin gene deletions. Cell **20**, 119-130.

Lerner, M. R., Boyle, J. A., Mount, S. N., Wolin, S. L. and Steitz, J. A. (1980). Are snRNPs involved in splicing? Nature **283**, 220-224.

Leder, P., Hansen, J. N., Konkel, D., Leder, A., Nishioka, Y. and Talkington, C. (1980). Mouse globin system: A functional and evolutionary analysis. Science **209**, 1336-1342.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eukaryotic DNA. Cell **15**, 687-701.

Mathis, D. and Chambon, P. (1981). The SV40 early region TATA box is required for accurate in vitro initiation of transcription. Nature **290**, 310-315.

Maxam, A. M. and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. In Methods in Enzymology, 65, pp. 499-560.

McClelland, A., Smith, D. F. and Glover, D. N. (1981). Short intervening sequences close to the 5' ends of the three Drosophila larval serum protein I genes. J. Mol. Biol. **153**, 257-272.

McKnight, S. L., Gavis, E. R., Kingsbury, R. and Axel, R. (1981). Analysis of transcriptional regulatory signals of the HSV thymidine kinase gene: Identification of an upstream control region. Cell **25**, 385-398.

Meyerowitz, E. M. and Hogness, D. S. (1982). Molecular organization of a Drosophila puff site that responds to ecdysone. Cell **28**, 165-176.

Miller, J. H. and Reznikoff, W. S. (1980). The Operon. Cold Spring Harbor Laboratory.

Mount, S. M. and Steitz, J. A. (1981). Sequence of U1 RNA from Drosophila melanogaster: Implications for U1 secondary structure and possible involvement in splicing. Nucl. Acids Res. **9**, 6351-6368.

Myers, R. M., Rio, D. C., Robbins, A. K. and Tjian, R. (1981). SV40 gene expression is modulated by the cooperative binding of T antigen to DNA. Cell **25**, 373-384.

Nishioka, Y., Leder, A. and Leder, P. (1980). An unusual alpha globin-like gene that has cleanly lost both globin intervening sequences. Proc. Natl. Acad. Sci. USA **77**, 2806-2809.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature **263**, 211-214.

Proudfoot, N. J. and Maniatis, T. (1980). The structure of a human α-globin pseudogene and its relationship to a α-globin gene duplication. Cell **21**, 537-544.

Rao, R. N. and Rogers, S. G. (1979). Plasmid pKC7: a vector containing ten restriction endonuclease sites suitable for cloning DNA segments. Gene **7**, 79-82.

Riddell, D. C., Higgins, M. J., McMillan, B. J. and White, B. N. (1981). Structural analysis of the three vitellogenin genes in Drosophila melanogaster. Nucl. Acids Res. **9**, 1323-1338.

Rio, D., Robbins, A., Myers, R. and Tjian, R. (1980). Regulation of simian virus 40 early transcription in vitro by a purified tumor antigen. Proc. Nat. Acad. Sci. USA **77**, 5706-5710.

Rodgers, J. and Wall, R. (1980). A mechanism for RNA splicing. Proc. Natl. Acad. Sci. USA **77**, 1877-1879.

Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J. L., LeMeur, M., Brégégégre, F., Gannon, F., LePennec, J. P., Chambon, P. and Kourilsky, P. (1979). The ovalbumin gene region: Common features in the organization of 3 genes expressed in the chicken oviduct under hormonal control. Nature **279**, 125-132.

Sánchez, F., Natzle, J. E., Cleveland, D. W., Kirschner, M. W. and McCarthy, B. J. (1980). A dispersed multigene family encoding tubulin in Drosophila melanogaster. Cell **22**, 845-854.

Sánchez, F., Tobin, S. L., Rdest, U., Zulauf, E. and McCarthy, B. J. (1982). Transcriptional and structural characteristics of two Drosophila actin genes. Submitted.

Scheller, R. H., McAllister, L. B., Crain, W. R., Durica, D. S., Posakony, J. W., Thomas, T. L., Britten, R. J. and Davidson, E. H. (1981). Organization and expression of multiple actin genes in the sea urchin. Molec. Cell. Biol. **1**, 609-628.

Smith, D. F., McClelland, A., White, B. N., Addison, C. F. and Glover, D. M. (1981). The molecular cloning of a dispersed set of developmentally regulated genes which encode the major larval serum protein of D. melanogaster. Cell **23**, 441-449.

Smith, D. R. and Calvo, J. M. (1980). Nucleotide sequences of the E. coli gene coding for dihydrofolate reductase. Nucl. Acids Res. **8**, 2255-2274.

Snyder, M., Hirsh, J. and Davidson, N. (1981). The cuticle genes of Drosophila: A developmentally regulated gene cluster. Cell **25**, 165-177.

Spradling, A. C. (1981). The organization and amplification of two chromosomal domains containing Drosophila chorion genes. Cell **27**, 193-201.

Spradling, A. C., Digan, M. E., Mahowald, A. P., Scott, M. and Craig, E. A. (1980). Two clusters of genes for major chorion proteins of Drosophila melanogaster. Cell **19**, 905-914.

Steinmetz, M., Moore, K. W., Frelinger, J. G., Taylor-Sher, B., Shen, F.-W., Boyse, E. A. and Hood, L. (1981). A pseudogene homologous to mouse transplantation antigens: transplantation antigens are encoded by eight exons that correlate with protein domains. Cell **25**, 683-692.

Steinmetz, M., Winoto, A., Minard, K. and Hood, L. (1982). Clusters of genes encoding mouse transplantation antigens. Cell **28**, 489498.

Tobin, S. L., Zulauf, E., Sánchez, F., Craig, E. A. and McCarthy, B. J. (1980) Multiple actin-related sequences in the Drosophila genome. Cell **19**, 121-131.

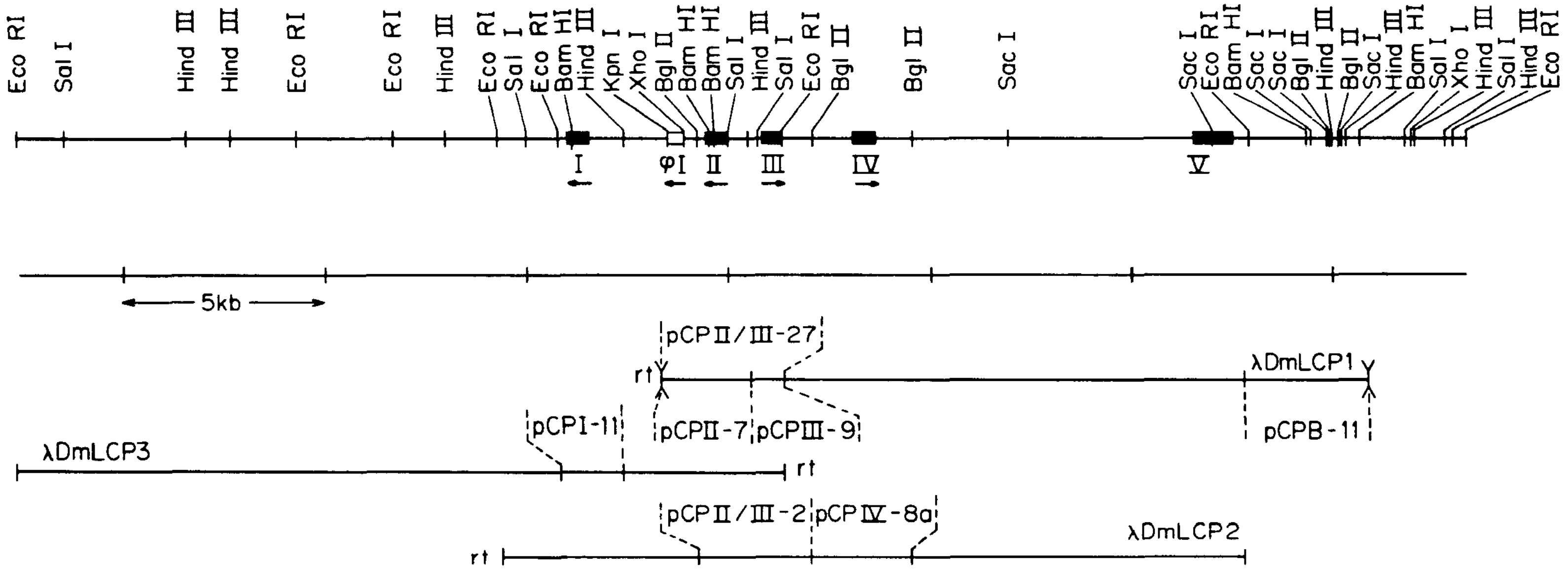
Vanin, E. F., Goldberg, G. I., Tucker, P. W. and Smithies, O. (1980). A mouse alpha globin-related pseudogene ($\psi\alpha$30-5) lacking intervening sequences. Nature **286**, 222-226.

Wasylyk, B. and Chambon, P. (1981). A T to A base substitution and small deletions in the conalbumin TATA box drastically decreased specific in vitro transcription. Nucl. Acids. Res. **9**, 1813-1824.

Wasylyk, B., Derbyshire, R., Guy, A., Molko, D., Roget, A., Teuole, R. and Chambon, P. (1980). Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in TATA box homology sequence. Proc. Natl. Acad. Sci. USA **77**, 7024-7028.

Figure 1. Restriction Map of the Cloned Cuticle Gene Region.

The upper diagram shows a composite restriction endonuclease map of the 44D cuticle gene region as derived from the lamba clone inserts depicted beneath it. Note that λDmLCP1 insert terminates with synthetic Eco RI sites while the inserts of λDmLCP2 and 3 contain natural EcoRI sites at their ends. This map was derived and coding regions for third instar messages localized as described in Snyder et al. (1981). To the left of λDmLCP1 only Hind III, Sal I, Kpn I and Eco RI sites are mapped. To the right of λDmLCP1 restriction sites were derived from a clone, λDmLCP13, which overlaps the right end of the map by 5.4 kb and extends an additional 13.7 kb further right (not shown). The arrow beneath each gene indicates the 5' to 3' direction of transcription. Subclones used in this study are derived as indicated.


Eco RI
Sal I
Hind III
Hind III
Eco RI
Eco RI
Hind III
Eco RI
Sal I
Eco RI
Bam HI
Hind III
Kpn I
Xho I
Bgl II
Bam HI
Bam HI
Sal I
Hind III
Sal I
Eco RI
Bgl II
Bgl II
Sac I
Sac I
Eco RI
Bam HI
Sac I
Sac I
Bgl II
Hind III
Bgl II
Sac I
Hind III
Bam HI
Sal I
Xho I
Hind III
Sal I
Hind III
Eco RI
I
φI
II
III
IV
V
5kb
pCPII/III-27
rt
λDmLCP1
pCPII-7
pCPIII-9
pCPB-11
pCPI-11
λDmLCP3
rt
pCPII/III-2
pCPIV-8a
λDmLCP2
rt

1

Figure 2. The Protein Coding Sequence of Genes I through IV.

The amino acid sequence of CP1 through CP4 is depicted directly over the encoding DNA sequence. The sequences are aligned according to their homology. A line indicates identity to gene I in that position. A gap indicated by [ ] was created in each of genes II through IV to give better sequence alignment. \/ indicates the position of a short intron as discussed in the text. The DNA sequence of genes I through IV was determined as described in Figure 7 and Experimental Procedures. The coordinate system above the map is based on the DNA sequence of gene I. The protein sequence was determined as described by Hunkapiller and Hood (1980). The amino acid residues determined from the N-terminus of in vivo proteins were 2-60 for CP1, 2-56 for CP2, 1-72 for CP3 and 1-64 for CP4. The remainder of the protein sequence is deduced from the DNA sequence. In genes I and II, an imperfect tandem duplication of 33 bp occurs at positions 22-54 and 55-87.

```
                                                        <-- Signal Peptide | In Vivo Protein -->
              INTRON
          -48      V                                                           1                               22                                          50
CP1       MetPheLysPheValMetIleCysAlaValLeuGlyLeuAlaValAlaAsnProProValProHisSerLeuGlyArgSerGluAspValHisAlaAsp
GENEI     ATGTTCAAGTTTGTCATGATCTGCGCAGTTTTGGGCCTGGCGGTGGCCAACCCCCCGGTGCCCCATTCCCTAGGCCGTTCGGAGGATGTCCACGCCGAT

                    V
CP2       Met-----------------------------Leu-----------Val------Val------Thr------[                ]LeuAlaProValSer-----------------Asp------------------------
GENEII    ATG-------------------G-------TCT-------C-----G------AG-------TACC-------[                ]-TAGCCC-AG-TTC------------C----C---T--------A--------T-----
                    V
CP3       Met-----------IleLeuLeuVal--------SerLeuAlaAla-----------ValAla--------[                                                    ]AsnAlaAsnValGlu
GENEIII   ATG-----------A-CC-GC-TG-----------TT-TC-CGCC-C-------------T---CC-------[                                                    ]A-C-CTA-T-TG-------G
                    V
CP4       Met-----------IleLeuLeuVal---------------LeuValAla--------ValAla--------[                                                    ]AsnGluAsnProGlu
GENEIV    ATG-----------A-CC-GC-TG----------------CC----G-C-C-----------T---CC-------[                                                    ]A-C-AGA-TC---------G

           54/55                                                 87                  100                                                                           150
CP1       ValLeuSerArgSerAspAspValArgAlaAspGlyPheAspSerSerLeuHisThrSerAsnGlyIleGluGlnAlaAlaSerGlyAspAlaHisGly
GENEI     GTCCTTTCCCGATCCGATGATGTTCGTGCCGATGGATTCGATTCCAGCCTGCACACCTCCAACGGAATCGAGCAGGCCGCCAGCGGTGATGCCCATGGC

CP2       ------------------------------------------------------------------------------------------------------------------------------------------------------
GENEII    ----------------------G---C----C-----------------C-------------C--------------------------A-----------------------------------------------------------

CP3       -----LysGluLeuValAsn----------GlnPro-----------------Val-------Lys-----ValLeuAspAsp----------SerAlaSerSer----------Thr-----------Ile----------
GENEIII   -----AAGGAG-TGGT-A-C-------------C-AGC-----------------C---T-TCAG-------AGT----GT-CT-GA-G------------TCT-CCTCCT-----------C-----A---CAT--------C-----

CP4       -----LysGluLeuValAsn----------Gln----------------------Val-------Lys-----ValLeuAsp-------------SerAlaAlaSer----------Thr-----------Val----------
GENEIV    -----AAGGAA-TGGT-A-C-------------C-AP------------------C------TAAG--------AGT-AGT-CTGGA-------------TTC-----CTGCTT-T---------T-C-------A-------T-----C---A

                                                                                  200
CP1       AsnIleHisGlyAsnPheGlyTrpIleSerProGluGlyGluHisValGluValLysTyrValAlaAsnGluAsnGlyTyrGlnProSerGlyAlaTrp
GENEI     AACATCCACGGCAACTTCGGCTGGATCTCACCCGAGGGCGAGCACGTCGAGGTTAAGTACGTCGCCAATGAGAACGGATACCAGCCCTCGGGAGCCTGG

CP2       ------------------------------------------------------------------------------------------------------------------------------------------------------
GENEII    ------------------------------------------------------------------------T--------A------------------G--------A-----------------------------------------

CP3       ----------Asp------Val------Glu-----------------------------Val----------Arg------Ser------Lys------Asp-----------------------------Gln------AspLeu
GENEIII   ----------G------AGT----------AG--------------C-------------T-TC-----T----GCGA-----G-GC-------AAG-----TG-C--------------------------------CA-A-T-A-CT-

CP4       ----------Asp------Val------Glu------Val--------------------------------Arg------Ser------Lys------Asp-----------------------------Gln------AspLeu
GENEIV    ----------G------AGTT---------AG------G------C------------------A---------CGT-------G-GC-------AAG-------G-C------------------------------CA-A-C-A-CTC

          250                                                                     300                                                                       345
CP1       IleProThrProProProIleProGluAlaIleAlaArgAlaValAlaTrpLeuGluSerHisProProAlaProGluHisProArgHisHisTrm
GENEI     ATCCCCACTCCTCCTCCAATCCCAGAGGCCATCGCCCGCGCCGTCGCCTGGCTAGAGTCCCACCCACCAGCACCCGAGCACCCCCGTCATCACTAG

CP2       ---------------------------------------------------------------------------------------------------------------------------------------------Trm
GENEII    --------------------------------------------------------------------T------------G--------T---------C---------------------------------------------

CP3       Leu----------------------------------Ala-----------LeuLys--------Ile--------TyrIle-----------AlaAsn--------SerLysAsnTrm
GENEIII   C-G-----------------------G------------CT-----------CTGAAG----------TA-----------ACA-C-----------G-TA-----------CAGCAAGAA-T-A

CP4       Leu------------------------------------------------LeuLys--------Ile--------TyrIleGlnAla--------------SerLysGluTrm
GENEIV    C-G------------------------------------------------CTGAAG-----------A-----------ACA-CC----------G-------T---CAGCAAGGAAT-A
```

Figure 3. Sequence Comparisons of Introns and of the 5' End of mRNAs for Drosophila Genes.

(a) A cDNA was synthesized from a labeled DNA primer and analyzed on 80 cm sequencing gels. Lanes a and c: cDNA to gene I DNA primer; lanes b and d are an overexposure of a sequencing T + C lane used here as a molecular length standard. Samples a and b were subjected to electrophoresis for a longer period than the samples on lanes c and d. A similar experiment was performed using a gene III DNA primer (not shown). The resulting cDNAs were sequenced to give the results presented in panels B and C. (b) Intron sequences in these four Drosophila cuticle genes. All of the intron-exon junctions are aligned as indicated. The gaps are included only to align the intron ends. Homology is indicated by boxes. For comparison the intron sequences of actin 88F and the sequences of intron/exon boundaries of other Drosophila genes are also displayed. Actin 88F and 79B (top) (Sánchez et al., submitted). Actin 79B (bottom) 57A and 5C (Fyrberg et al., 1981). YP1 = yolk protein gene I (Hovemann et al., 1981). 83K HSP = 83 kilodalton heat shock protein (Holmgren et al., 1981). ADH = alcohol dehydrogenase (Benjayati et al., 1980). 68C = genes from 68C salivary gland puff site (M. D. Garfinkel, R. E. Pruitt and E. M. Meyerowitz, in preparation). Dm U1 RNA = appropriate region from Drosophila U1 RNA (Mount and Steitz, 1981) that is complementary to intron boundary sequences (Lerner et al., 1980; Rogers and Wall, 1980). Eucaryotic consensus sequence (Lerner et al., 1980). The Drosophila consensus derived from these data is such that a large letter represents a base present in greater than 2/3 of the cases shown. Smaller letters indicate preferred alternative bases. P = Purine, Y = Pyrimidine. (c) The nucleotide sequence at the 5' end of the mRNA. Homology to gene I is indicated by a horizontal line and gaps [] are included for better sequence alignment. Homologies between genes III and IV are indicated by vertical lines in the regions which are non-homologous with gene I. (continues)

Figure 3 (continued)

The * indicates the length of the principal cDNA synthesized from labeled DNA primer (see text). The 5' end of the primer used is depicted. The ↓ indicates the probable first nucleotide of the mRNA as discussed in the text. The position of introns is also indicated. Codons are shown by an underline.

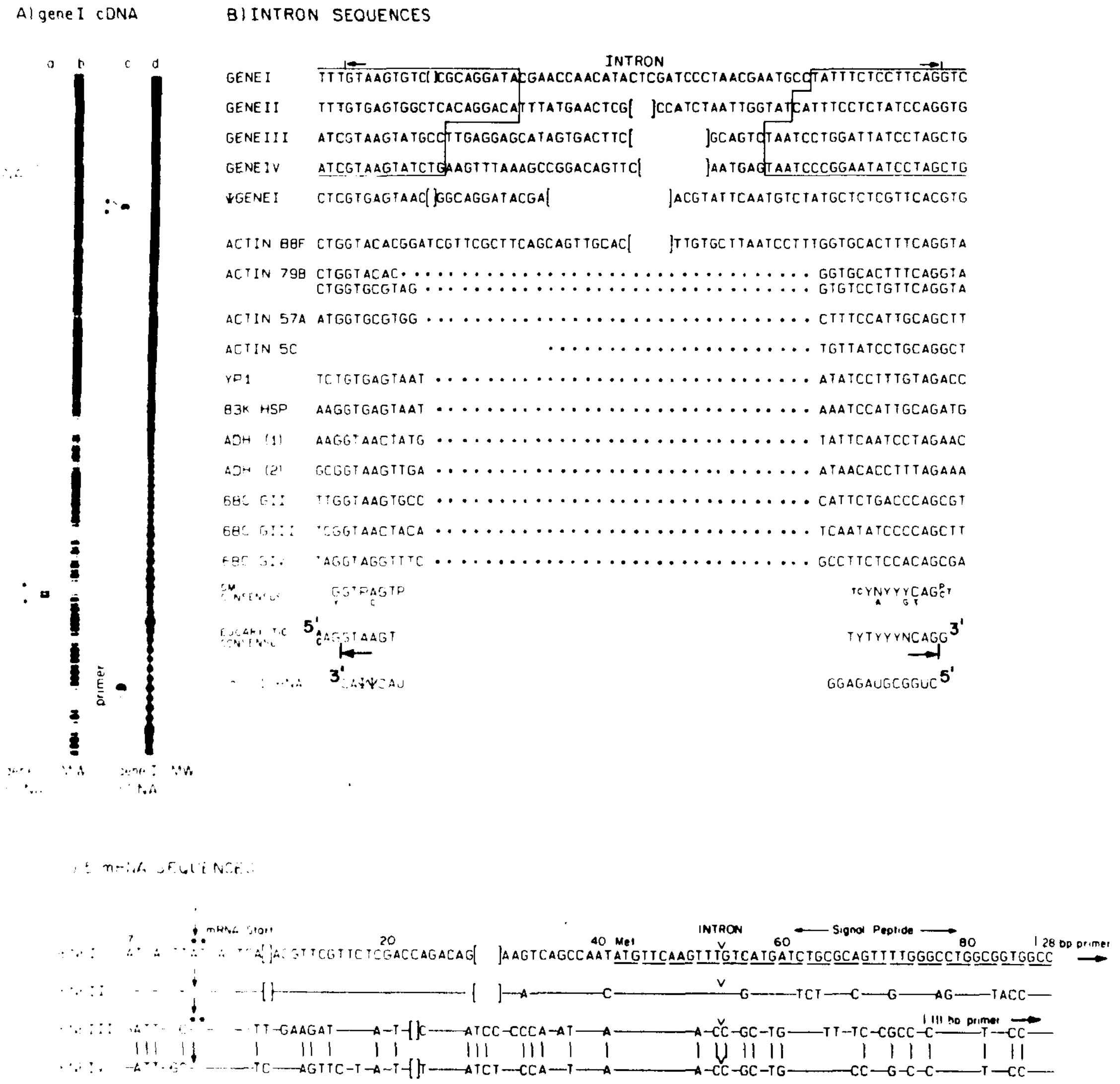

A) gene I cDNA
a
b
c
d
primer
B) INTRON SEQUENCES
INTRON
GENE I TTTGTAAGTGTC[ ]CGCAGGATACGAACCAACATACTCGATCCCTAACGAATGCCTATTTCTCCTTCAGGTC
GENE II TTTGTGAGTGGCTCACAGGACATTTATGAACTCG[ ]CCATCTAATTGGTATCATTTCCTCTATCCAGGTG
GENE III ATCGTAAGTATGCCTTGAGGAGCATAGTGACTTC[ ]GCAGTCTAATCCTGGATTATCCTAGCTG
GENE IV ATCGTAAGTATCTGAAGTTTAAAGCCGGACAGTTC[ ]AATGAGTAATCCCGGAATATCCTAGCTG
ΨGENE I CTCGTGAGTAAC[ ]GGCAGGATACGA[ ]ACGTATTCAATGTCTATGCTCTCGTTCACGTG
ACTIN 88F CTGGTACACGGATCGTTCGCTTCAGCAGTTGCAC[ ]TTGTGCTTAATCCTTTGGTGCACTTTCAGGTA
ACTIN 79B CTGGTACAC.............................. GGTGCACTTTCAGGTA
CTGGTGCGTAG.............................. GTGTCCTGTTCAGGTA
ACTIN 57A ATGGTGCGTGG.............................. CTTTCCATTGCAGCTT
ACTIN 5C .................... TGTTATCCTGCAGGCT
YP1 TCTGTGAGTAAT.............................. ATATCCTTTGTAGACC
83K HSP AAGGTGAGTAAT.............................. AAATCCATTGCAGATG
ADH (1) AAGGTAACTATG.............................. TATTCAATCCTAGAAC
ADH (2) GCGGTAAGTTGA.............................. ATAACACCTTTAGAAA
68C GII TTGGTAAGTGCC.............................. CATTCTGACCCAGCGT
68C GIII TCGGTAACTACA.............................. TCAATATCCCCAGCTT
68C GIV TAGGTAGGTTTC.............................. GCCTTCTCCACAGCGA
SM CONSENSUS GGTRAGTP TCYNYYYCAGPCT
EUCARYOTIC CONSENSUS 5' AAGGTAAGT TYTYYYNCAGG 3'
U1 RNA 3' UAΨΨCAU GGAGAUGCGGUC 5'
C) 5' mRNA SEQUENCES
mRNA Start
7
20
40 Met
INTRON
60
Signal Peptide
80
28 bp primer
GENE I A[ ]ACGTTCGTTCTCGACCAGACAG[ ]AAGTCAGCCAATATGTTCAAGTTTGTCATGATCTGCGCAGTTTTGGGCCTGGCGGTGGCC
GENE II [ ]——[ ]—A——C——G——TCT—C—G—AG—TACC—
GENE III TT-GAAGAT—A-T-[ ]C—ATCC-CCCA-AT—A—A-CC-GC-TG—TT-TC-CGCC-C—T-CC—
111 bp primer
GENE IV TC—AGTTC-T-A-T-[ ]T—ATCT—CCA—T—A—A-CC-GC-TG—CC—G-C-C—T—CC—

Figure 4. Sequence Comparison of the Regions Flanking the Cuticle Genes. (a) Homology in the upstream flanking sequence. Homologous regions shared by all four genes are depicted by dark boxes. Sequences shared by only two genes are depicted in light boxes. A repeated sequence in gene IV is indicated by arrows. Palindromes are indicated by 〰〰 underline. (b) The 3' untranslated region. Poly(A) addition consensus sequences, AATAAA, for genes III and IV or similar ones for genes I and II are indicated by small leters.

A) 5' FLANKING REGION

```
                                                                     -250                                             -200
GENEI    ATGAAGGGGCGGGTCAAACTCGTTTCGATTTCGGGATGCCACCCGACCCGTTTGCCCCTTTATTGATGCGATTGTTTCATTTTAGCATCTATTAAGCGATT
GENEII   TTGGTCAACTCGAAACGTCATCGATGAGCACAGAATCCGAAAACCGTACTCCATCGCCCCTACAAAATTTCTACCGAAGCATGTTTCATTTCGGAATCTG
GENEIII  AAGTTTAGTTCATATTGAAGTTGAATTTTAGAAAATTAAATATTGTACTGCTTAATAATTATTCTGGTTTCTGGTCCGGTTTGCTTTGCATTTCGGTTAG
GENEIV   AAGATGTTAAATTATTTCAATGAGAAGGTGGACTTACCCTTTTCCGAGTAACCCGATTCTTTTTAGAATAATTACGGTAGCGATTTGCATAGACAATAGA

                                                                     -150                                             -100
GENEI    ATATATAGTACTTATCCCGTTGTTTGGCATTTGCTAAGCTGTCGCATGTGACGATGCTTTTTAATGGGTGTGGGCGCATCCGCGAAGTCAACCCATAACT
GENEII   TTCAGCAGCGCAAGACTTGTTTTTTGACATTTGTATCGCAGAGTCAAGTGGAGAATTTATGGGCCCTGCCTTTTGTTGGCATCATGGGCGTTTCGTGATA
GENEIII  ACTAGGGCGAATATTTCAGTTGAATAAATAACTAAGAATGCTCATCTCCTAATGAAAGTGGTTAAGCCATCTCAAGTCGACTAATTTGCATCCCAGACGG
GENEIV   AATCAAAAAGAGTGCAGCAGACGATTTTTATCGCCACCAAGCATGTCACTTGAACCAGTCCGTAAAACCAAACGAGACCTATGCTGGCCGAAATGTTAAT

                                                                     -50                   TATA BOX        C          -1 ↓mRNA Start
GENEI    CAGCGAACCAATTGAATGCAAGATGTAGAGTTTTGATATGGGTTCACTTTGGGTGGCAATCATATAAAAAGGCTCTGCCGACCACAATCAGTTATCAGTC
GENEII   ACTTAGATTTGGCCCAAAAAGTAATAAGCAATTCGTTTGGAAAGCAACCAAATTGGGAATCATATAAAAAGACTCTGTCGACCAAAGTCAGTTATCAGTC
GENEIII  TTTTTATTATATGCATCACATTGACTTAATTATAATACGCACATTGCATCAGCTTTTGATGATATATAAACACCGATTTGAGCATAGATTGTCATCAGTC
GENEIV   TAAAAACGGGTTGCATCAGCTTTTGATCAGCTTTAAGATTTCGTGGGGAGTGCGTATAAGCGTATAAAAGCCGACGAGTGATCCCGAATTGGCATCAGTC
```

B) 3' UNTRANSLATED REGION

```
         Trm1                                            50
GENEI    TAGAACCTCTATGAAAGCGGATCGCACTACGGACTGTTCCCCGAAGACCTTTCGAACTATTAGCTTAAGTAATCGTACTGTTTGTAAaatacaCGCAATT
GENEII   TAGGACTCGTCACCCGGATCCCGGACCACTACACGGACTGTTCTCCCGAAACAAATCGCCCAAGTTGTTTAGCTGTACTTCTTGACTTTCAaaaaaaTAC
GENEIII  TAAGTGAACCCGCCGACTAGGAACATGAAAGATTGGAGACAGCTAGGTTGAGTTTGGATAATTTCTTACCAGTTGTTTTAAATTTAAGGAAAATGTTATC
GENEIV   TAAGCAATCGACACGACCAGGACCCACATTCGAATCGGAGGTGCAACTCCAAAGACCTTGCCCTCTAACCCTTAGAATTTAAACAGCATGCAGACATTAT
            100                                          150
GENEI    GTTAACGGCAGAAACCAGTTTGCAACCTTGACTTTGAATTTGGCAAACAACTGTAACGGTTTCGAACCCGTCCTACCCGTTTACCACCTTCGATTTACTT
GENEII   ATGCACTTGCTTATAGCagtaaaAATGTGTGTGTTCTTCATTGCACTTTTTAGGTAGTTCCTGTAATAATACGAGCTTTTATACCTCTACCTTCGCTGGG
GENEIII  GAATTCGAAaataaaTTAAACCTTGCAATATAAACCAAGTGCATGTTTTACAAATCTGACAGTTCGATTTAAGAGAAGGCTCCCGGTATTATATGGTATA
GENEIV   AAATGATTATCGAGTTAGGAaataaaTTCGATACTCTTTGGCAACAAATCTATTTAGATATGGTCTTAATTTCCCTGACGGCAGCAGGAGTTACCTTGTT
```

Figure 5. Sequence Comparisons of ψ Gene I with Genes I and II

(a) A best fit comparison between a 1.70 kb DNA sequence containing gene I and a 2.26 kb DNA fragment containing genes II and ψ gene I calculated according to Hunkapiller and Hood (in preparation). The comparison is such that every 40 nucleotides of the gene I fragment is compared to the gene II/ψ gene I sequence. A diagonal line therefore indicates homology. A schematic representation of each gene is depicted on the axes, and units are given in 100 bp. ii) A comparison between the gene II and ψ gene I sequences. Gene II is aligned directly beneath gene I in panel i). (b) Nucleotide sequence comparisons of genes I, II and ψI. Lines indicate identity with gene I, [ ] indicate gaps inserted for sequence alignment. Translation of ψ gene I from position 117 to 468 is in its single long open reading frame.

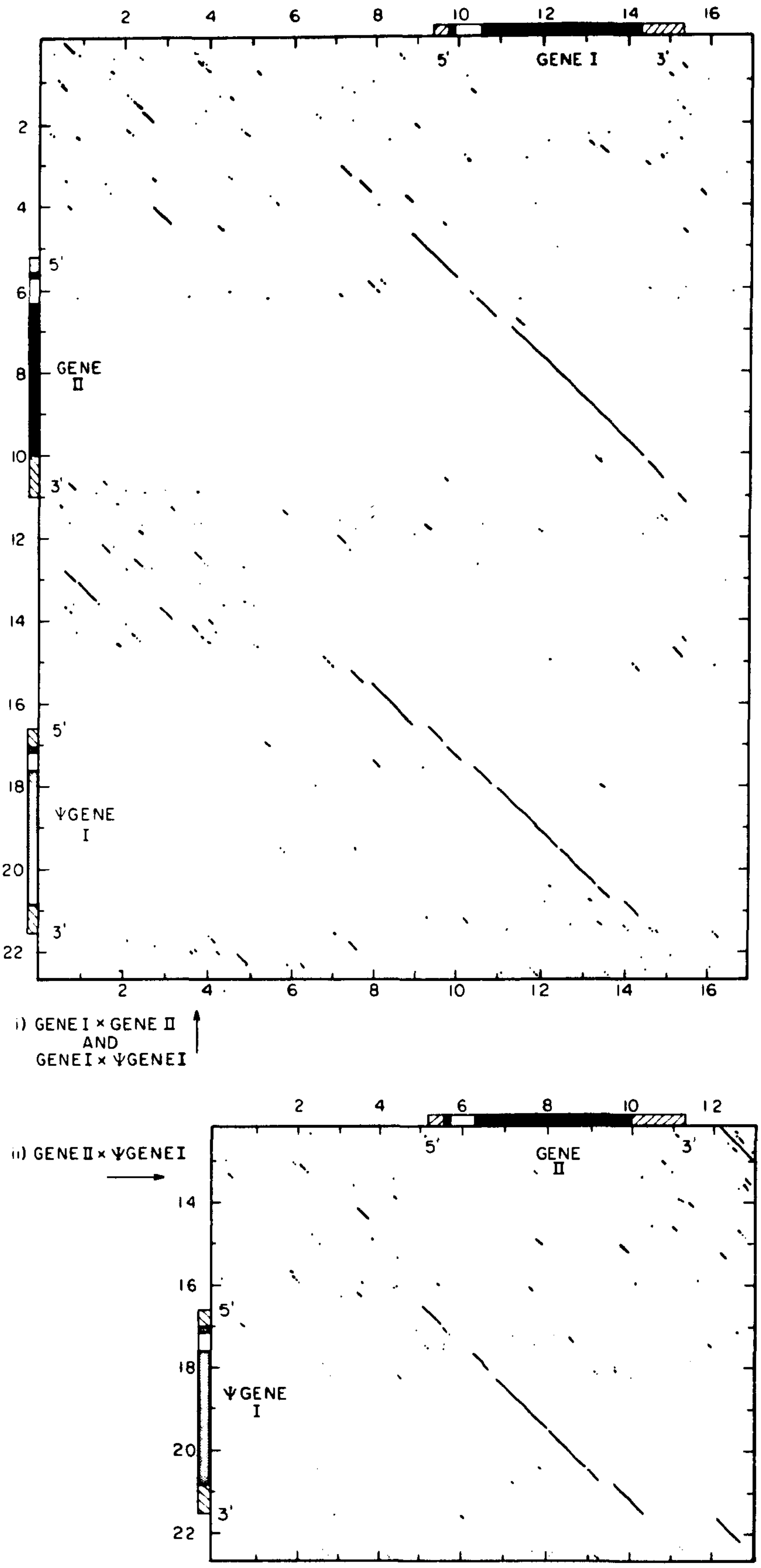
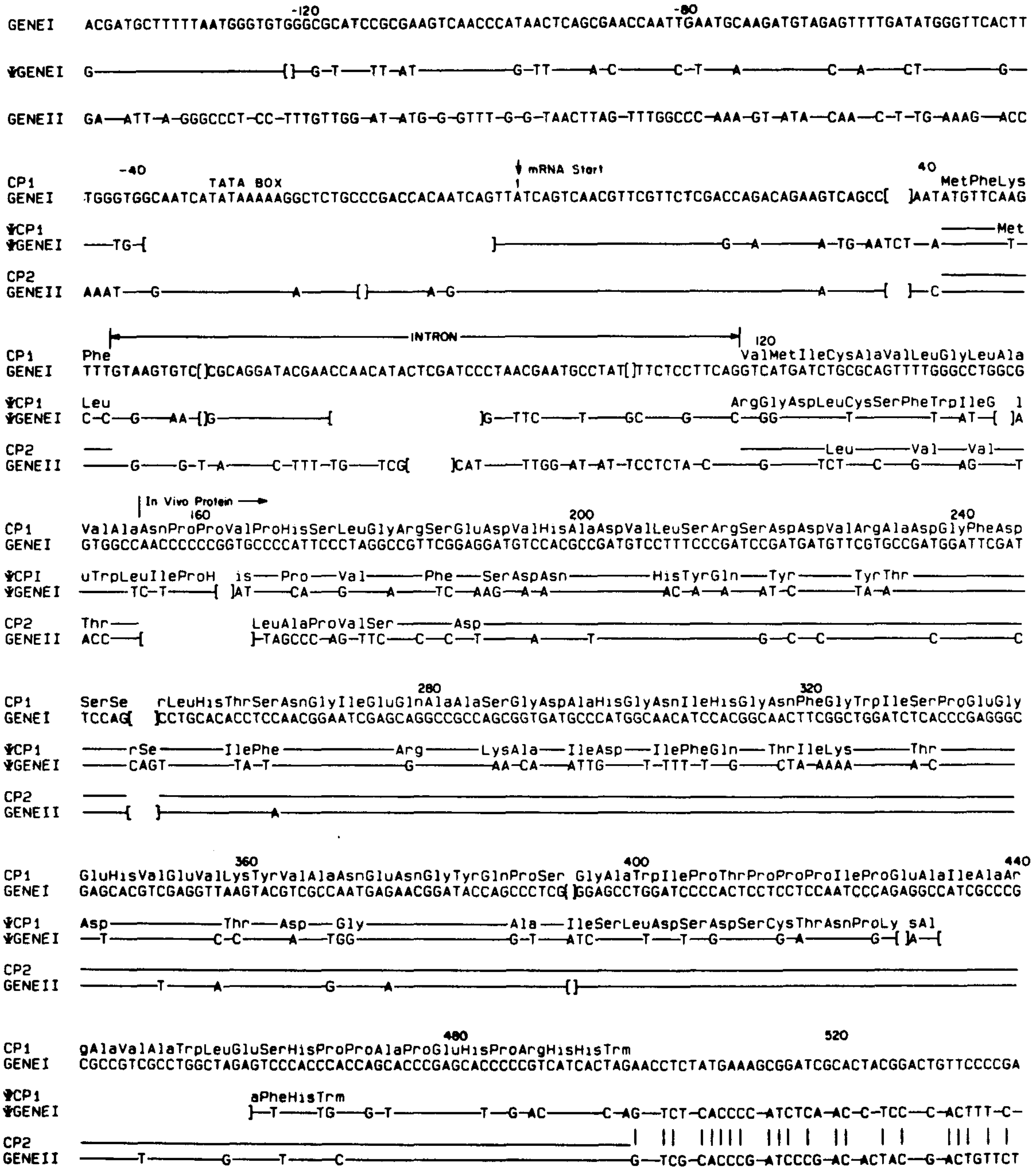


**Figure 5 (left)**



-120
-80
GENEI ACGATGCTTTTTAATGGGTGTGGGCGCATCCGCGAAGTCAACCCATAACTCAGCGAACCAATTGAATGCAAGATGTAGAGTTTTGATATGGGTTCACTT
ΨGENEI G—{}—G-T—TT-AT—G-TT—A-C—C-T—A—C—A—CT—G—
GENEII GA—ATT-A-GGGCCCT-CC-TTTGTTGG-AT-ATG-G-GTTT-G-G-TAACTTAG-TTTGGCCC-AAA-GT-ATA—CAA—C-T-TG-AAAG—ACC
-40
TATA BOX
mRNA Start
1
40
CP1 MetPheLys
GENEI TGGGTGGCAATCATATAAAAAGGCTCTGCCCGACCACAATCAGTTATCAGTCAACGTTCGTTCTCGACCAGACAGAAGTCAGCC[ ]AATATGTTCAAG
ΨCP1 —Met
ΨGENEI —TG-{ }—G—A—A-TG-AATCT—A—T-
CP2
GENEII AAAT—G—A—{}—A-G—A—{ }—C—
INTRON
120
CP1 Phe ValMetIleCysAlaValLeuGlyLeuAla
GENEI TTTGTAAGTGTC[]CGCAGGATACGAACCAACATACTCGATCCCTAACGAATGCCTAT[]TTCTCCTTCAGGTCATGATCTGCGCAGTTTTGGGCCTGGCG
ΨCP1 Leu ArgGlyAspLeuCysSerPheTrpIleG l
ΨGENEI C-C—G—AA-{]G—{ ]G—TTC—T—GC—G—C—GG—T—T—AT—{ ]A
CP2 —Leu—Val—Val—
GENEII —G—G-T-A—C-TTT-TG—TCG{ ]CAT—TTGG-AT-AT-TCCTCTA-C—G—TCT—C—G—AG—T
In Vivo Protein
160
200
240
CP1 ValAlaAsnProProValProHisSerLeuGlyArgSerGluAspValHisAlaAspValLeuSerArgSerAspAspValArgAlaAspGlyPheAsp
GENEI GTGGCCAACCCCCCGGTGCCCCATTCCCTAGGCCGTTCGGAGGATGTCCACGCCGATGTCCTTTCCCGATCCGATGATGTTCGTGCCGATGGATTCGAT
ΨCPI uTrpLeuIleProH is—Pro—Val—Phe—SerAspAsn—HisTyrGln—Tyr—TyrThr—
ΨGENEI —TC-T—{ ]AT—CA—G—A—TC—AAG—A-A—AC-A—A—AT-C—TA-A—
CP2 Thr— LeuAlaProValSer—Asp—
GENEII ACC—{ }-TAGCCC-AG-TTC—C—C—T—A—T—G—C—C—C—C
280
320
CP1 SerSe rLeuHisThrSerAsnGlyIleGluGlnAlaAlaSerGlyAspAlaHisGlyAsnIleHisGlyAsnPheGlyTrpIleSerProGluGly
GENEI TCCAG{ ]CCTGCACACCTCCAACGGAATCGAGCAGGCCGCCAGCGGTGATGCCCATGGCAACATCCACGGCAACTTCGGCTGGATCTCACCCGAGGGC
ΨCP1 —rSe—IlePhe—Arg—LysAla—IleAsp—IlePheGln—ThrIleLys—Thr—
ΨGENEI —CAGT—TA-T—G—AA-CA—ATTG—T-TTT-T—G—CTA-AAAA—A-C—
CP2 — —
GENEII —{ }—A—
360
400
440
CP1 GluHisValGluValLysTyrValAlaAsnGluAsnGlyTyrGlnProSer GlyAlaTrpIleProThrProProProIleProGluAlaIleAlaAr
GENEI GAGCACGTCGAGGTTAAGTACGTCGCCAATGAGAACGGATACCAGCCCTCG{ ]GGAGCCTGGATCCCCACTCCTCCTCCAATCCCAGAGGCCATCGCCCG
ΨCP1 Asp—Thr—Asp—Gly—Ala—IleSerLeuAspSerAspSerCysThrAsnProLy sAl
ΨGENEI —T—C-C—A—TGG—G-T—ATC—T—T—G—G-A—G—{ ]A—{
CP2 —
GENEII —T—A—G—A—{}—
480
520
CP1 gAlaValAlaTrpLeuGluSerHisProProAlaProGluHisProArgHisHisTrm
GENEI CGCCGTCGCCTGGCTAGAGTCCCACCCACCAGCACCCGAGCACCCCCGTCATCACTAGAACCTCTATGAAAGCGGATCGCACTACGGACTGTTCCCCGA
ΨCP1 aPheHisTrm
ΨGENEI }—T—TG—G-T—T—G-AC—C-AG—TCT-CACCCC-ATCTCA-AC-C-TCC—C-ACTTT-C-
CP2 —| || ||||| ||| | || | | ||| | |
GENEII —T—G—T—C—G—TCG-CACCCG-ATCCCG-AC-ACTAC—G-ACTGTTCT

Figure 5 (right)

Figure 6. Evolutionary Scheme for the Cuticle Gene Cluster.

A simple model for how the cuticle gene cluster evolved as described in the text. Several variations in the order of events from that shown in the figure are also plausible; in particular the inversion of genes I and II or their precursor, relative to III and IV or their precursor could occur after the gene segment duplication within a gene I/II precursor.

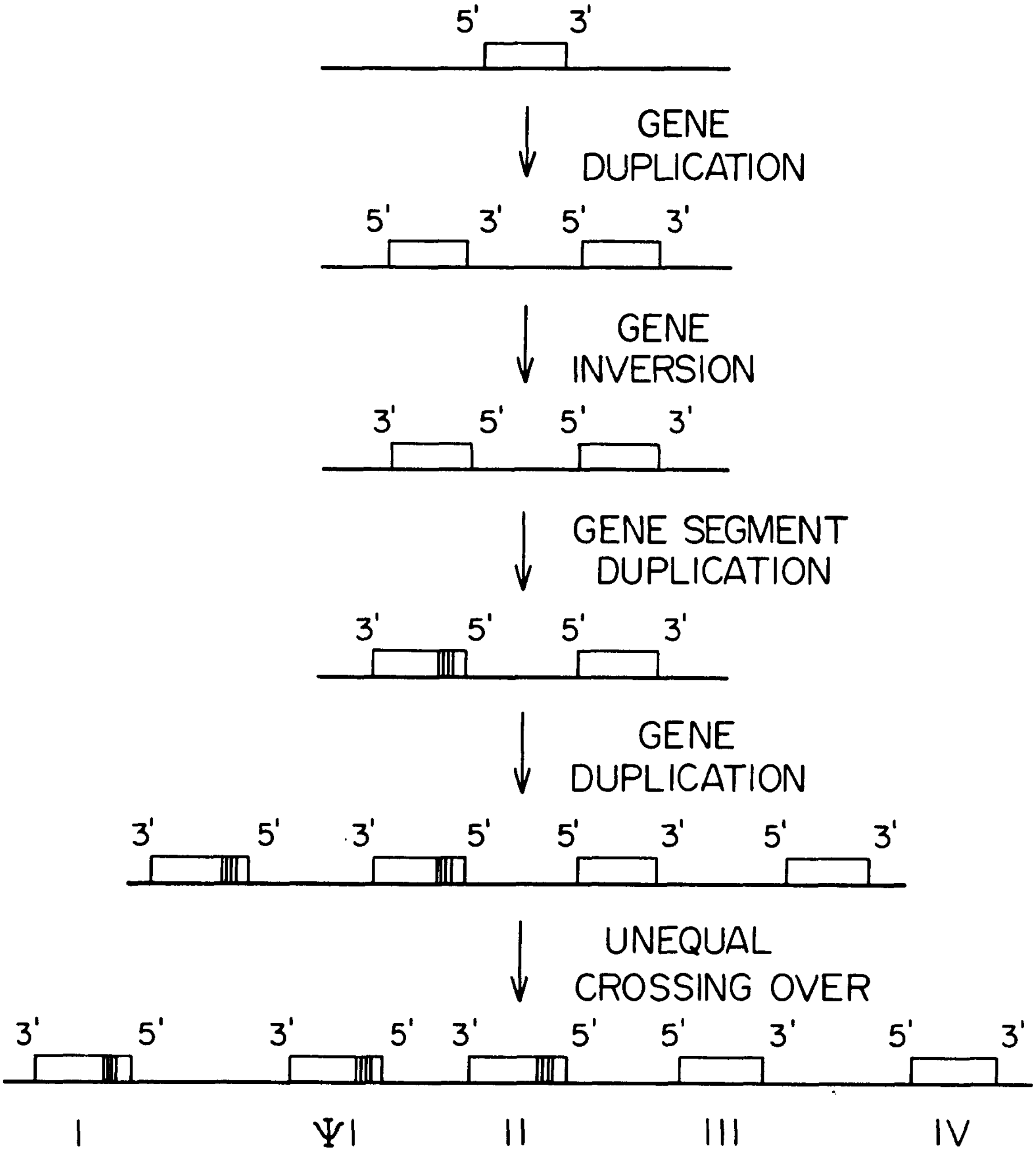

5'
3'
GENE
DUPLICATION
5'
3'
5'
3'
GENE
INVERSION
3'
5'
5'
3'
GENE SEGMENT
DUPLICATION
3'
5'
5'
3'
GENE
DUPLICATION
3'
5'
3'
5'
5'
3'
5'
3'
UNEQUAL
CROSSING OVER
3'
5'
3'
5'
3'
5'
5'
3'
5'
3'
I
ΨI
II
III
IV

Figure 7. Gene I through IV Restriction Map and Sequencing Scheme.

Restriction map of the region encoding genes I through IV is shown. All Hpa II, Ava I, Ava II sites are indicated, plus any other restriction sites pertinent to the sequencing strategy depicted beneath the map. Note that the sequenced regions of genes II, III and IV are continuous. The schematic representation of the genes is as follows: solid box = protein coding sequence; open box = introns; diagonally lined box = untranslated mRNA coding region; gray box = pseudogene protein coding. The DNA sequencing strategy is such that a vertical line (|) or circle (o) indicates $^{32}$P labeling at either a 3' or 5' terminus, respectively. The arrows indicate the direction and extent of sequencing from the labeled terminus. Dashed lines within several arrows indicate a region where the sequence was not read.

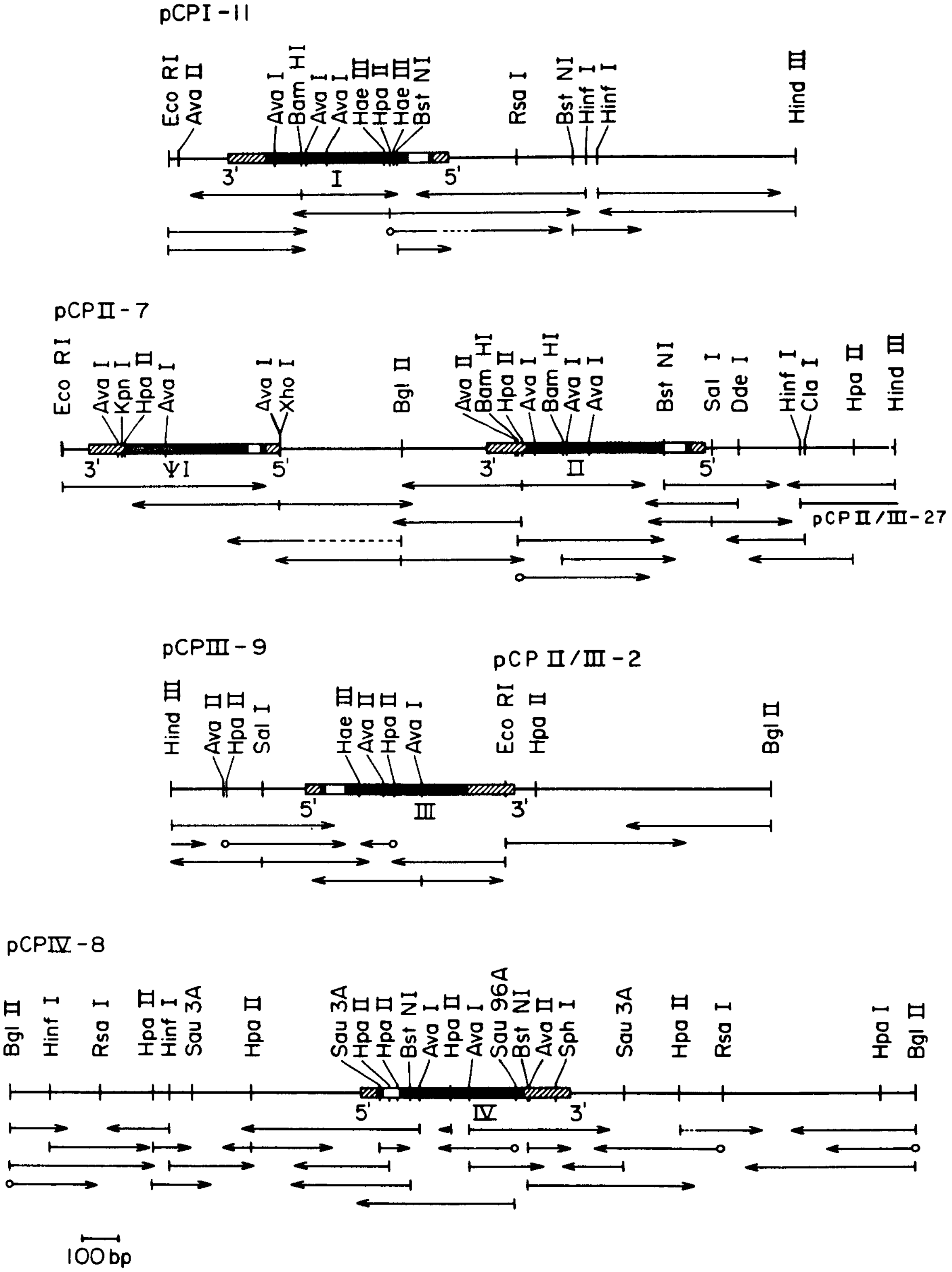

pCPI-11
Eco RI
Ava II
Ava I
Bam HI
Ava I
Ava I
Hae III
Hpa II
Hae III
Bst NI
Rsa I
Bst NI
Hinf I
Hinf I
Hind III
3'
I
5'
pCPII-7
Eco RI
Ava I
Kpn I
Hpa II
Ava I
Ava I
Xho I
Bgl II
Ava II
Bam HI
Hpa II
Ava I
Bam HI
Ava I
Ava I
Bst NI
Sal I
Dde I
Hinf I
Cla I
Hpa II
Hind III
3'
ΨI
5'
3'
II
5'
pCPII/III-27
pCPIII-9
Hind III
Ava II
Hpa II
Sal I
Hae III
Ava II
Hpa II
Ava I
5'
III
3'
pCPII/III-2
Eco RI
Hpa II
Bgl II
pCPIV-8
Bgl II
Hinf I
Rsa I
Hpa II
Hinf I
Sau 3A
Hpa II
Sau 3A
Hpa II
Hpa II
Bst NI
Ava I
Hpa II
Ava I
Sau 96A
Bst NI
Ava II
Sph I
Sau 3A
Hpa II
Rsa I
Hpa I
Bgl II
5'
IV
3'
100 bp

Figure 8. Appendix.

The entire DNA sequence determined from the Eco RI site 200 bp to the left of gene I to the Bgl II site 1 kb to the right of gene IV (see Figure 1) is presented in a 5' to 3' direction. Protein coding regions of genes I, II, III and IV are presented in capital letters; the remainder of the sequence is in small letters. A 0.8 kb gap in the I/II spacer is indicated. Also shown are restriction endonuclease sites pertinent to the subcloning. The boxed sequence at position 1697 is not present in the Drosophila genome, but was synthetically placed there the cloning procedures (Maniatis et al., 1978).

```
Eco RI
gaattcggtttacggtccaggtagagaccacattaccatattgtactgaacgtgctaaacaactaagtaaatcgaaggtggtaaacgggtaggacgggttcgaaaccgttacagttgtttgccaaattcaaagtcaaggttgcaaactgg 150
                                                                                                               Trm
tttctgccgttaacaattgcgtgtattttacaaacagtacgattacttaagctaatagttcgaaaggtcttcggggaacagtccgtagtgcgatccgctttcatagaggttCTAGTGATGACGGGGGTGCTCGGGTGCTGGTGGGTGGGA 300
                                                                          GENE I
CTCTAGCCAGGCGACGGCGCGGGCGATGGCCTCTGGGATTGGAGGAGGAGTGGGGATCCAGGCTCCCGAGGGCTGGTATCCGTTCTCATTGGCGACGTACTTAACCTCGACGTGCTCGCCCTCGGGTGAGATCCAGCCGAAGTTGCCGTG 450
GATGTTGCCATGGGCATCACCGCTGGCGGCCTGCTCGATTCCGTTGGAGGTGTGCAGGCTGGAATCGAATCCATCGGCACGAACATCATCGGATCGGGAAAGGACATCGGCGTGGACATCCTCCGAACGGCCTAGGGAATGGGGCACCGG 600
                                                                          INTRON I                                Start
GGGGTTGGCCACCGCCAGGCCCAAAACTGCGCAGATCATGACctgaaggagaaataggcattcgttagggatcgagtatgttggttcgtatcctgcggacacttacAAACTTGAACATattggctgacttctgtctggtcgagaacgaac 750
gttgactgataactgattgtggtcgggcagagcctttttatatgattgccacccaaagtgaacccatatcaaaactctacatcttgcattcaattggttcgctgagttatgggttgacttcgcggatgcgcccacacccattaaaaagca 900
tcgtcacatgcgacagcttagcaaatgccaaacaacgggataagtactatatataatcgcttaatagatgctaaaatgaaacaatcgcatcaataaggggcaaacgggtcgggtggcatcccgaaatcgaaacgagtttgacccgcccct 1050
tcattaaatcccctcctttgagcttgacttggcattgccaggccgtgaaccacttttgtagtcccatcaagactccatggactacaaagcatcgccgacttcttagactcgatgtttgttatatataaacgtgtgtctctaccttggca 1200
ccatttcttctttttttttatcaattcctctcgacgaacaggttgcagtttgcttttaaacaaattcaagtgtgttttacattatatttcatatgcatagcgccttatattattttttttatatttcgtaaaattaaataagataataagac 1350
ctgttcaatataggaggctaccgaaataaatacagtttttagcgacaattaaagttgttaagaaaacctaatgttaacctaaagaagatatttacttgtctgtataggcatataaactaaaagtttacttaaaatttataatticctata 1500
gatacatttcatggtatttaaatgtgcaagccgaacatctaatataagagcttgaatagtttgtttttcagatcagttagaccgcacagtatcctcacaaatttggtttgaaagttaagattaatttgtaatcaaacagaatagcctctt 1650
                                          Hind III     Eco RI
ctgagatttatgctagtgactatcgcacctttaacaatccaagctt/86Kb/gaattcatggtggaatgccgttcatggtgattcttctctctttttgaaaccatttacttatgtcgggtaaaaaactttattgaataatatataacatgt 1975
                                                            Trm
cgatggaaagtcggtggatgggtctgagatcggggtgaagagtcttggtgatggtgcggatgctcgggtaccggtcagtggaacgccttcggattggtgcaggagtcggaatcaaggctgatcgaagcctggtatccgttctcaccagcg 1945
tcgtacgtgacctcgacgtgatcgccctcggggggtgatccatttatagtgccctgaaaaataccatcaatatctgctttggcggcccgctcgattccgttgaatatgtgcagactgctggaatcgaatccatcggtataaacatcgtat 2095
                                                                         ↓GENE I
gattggtagtggacatcggcgttgtcactttcgaaacgtcctacggatgggggatgggggatgagccactcaatccaaaaactgcacagatcaccacgtgaacgagagcatagacattgaatacgttcgtatcctgccgttactcacgag 2245
    Start
catgaacatttagattccaatttctgtttgctcgagaacgaacgttgactgataaccaccaaaactgaacccaagtcaatacgctacatctttcatacgattggtgctctgaaatctgggttgactatgaagataccccacacccatta 2395
aaaagcatcgcatggcaatccacgggattagtactacgagtatatactagggtaatcggctcgtgtggcattgaccattgctttctatttcaataagtttgtccattggtattggcttctctaattggctttaaatactacttaaaat 2545
aaaaaaaaaatgcattgagactttagtgagaataaaattgtatgcaaacatttgctttttgtttttcagatcttttagactgcccagatccttctcaaatttaaaagacaaattaatttttaattcagcacatctttctttatgg 2695
acacatatactcatttttgataaatggatttagtgaatcgaatataaaggtagaaggaagcattcccagcgaaggtagaggtataaagctcgtattattacaggaactacctaaaaagtgcaatgaagaacacacacatttttactgc 2845
                                                                                                               Trm
tataagcaagtgcatgtatttttttgaaagtcaagaagtacagctaaacaacttgggcgatttgtttcgggagaacagtccgtgtagtggtccgggatccgggtgacgagtcCTAGTGATGACGGGGGTGCTCGGGTGCTGGGGGGTGAG 2995
                                                                          GENE II
ACTCCAGCCAGGCAACGGCGCGGGCGATGGCCTCTGGGATTGGAGGAGGAGTGGGGATCCAGGCTCCCGAGGGCTGGTATCCGTTTTCATTCGCGACGTACTTTACCTCAACGTGCTCGCCCTCGGGTGAGATCCAGCCGAAGTTGCCGT 3145
GGATGTTGCCATGGGCATCACCGCTGGCGGCCTGCTCGATTCCGTTTGAGGTGTGCAGGCTGGAGTCGAATCCGTCGGCACGAACGTCGTCCGATCGGGAAAGGACATCAGCGTGTACATCATCGGAGCGGGAAACTGGGGCTAGGGCGG 3295
                                                                         INTRON II                   Start
TAGCCACTCCCACAACGGCGAGAATCATCACctggatagaggaaatgataccaattagatggcgagttcataaatgtcctgtgagccactcacAAACTTGAACATgttggctgatttctgtctggtcgagaacgaacgttgactgataac 3445
tgactttggtcgacagagtctttttatatgattcccaatttggttgctttccaaacgaattgcttattactttttgggccaaatctaagttatcacgaaacgcccatgatgccaacaaaggcagggcccataaattctccacttgactc 3595
tgcgatacaaatgtcaaaaaacaagtcttgcgctgctgaacagattccgaaatgaaacatgcttcggtagaaattttgtaggggcgatggagtacggttttcggattctgtgctcatcgatgacgtttcgagttgaccaagtctttatca 3745
atcatgcctgtacaaaatcaaaaacccaatacacttcacttgatagcagtcccatctatttaaaaggttccgaagtggtaataaacaatataccggaataaacgatttcgcaagtgatactgcatattaacgtgctagttgcctatga 3895
                                                       HindIII
cattttgttgtatctcaatattttggatgtcactgtttaatttaaaaaacccgcaaagcttttggccagagttgtcaacgtgccacacaccaaatgaaacaccgaaaactatgctatgcttaagtttagttcatattgaagttgaattt 4045
tagaaaattaaatattgtactgcttaataattattctggtttctggtccggtttgctttgcatttcggttagactagggcgaatatttcagttgaataaataactaagaatgctcatctcctaatgaaagtggttaagccatctcaagtc 4195
gactaatttgcatcccagacggttttattatatgcatcacattgacttaattataatacgcacattgcatcagcttttgatgatatataaacaccgatttgagcatagattgtcatcagtcttagaagatttctagtccgacaatccac 4345
        Start                                                             INTRON III
ccaaatcaaaATGTTCAAGATCgtaagtatgccttgaggagcatagtgacttcgcagtctaatcctggattatcctagCTGCTTGTCTGTTCTCTCGCCGCCCTGGTGGCCGCCAACGCTAATGTGGAGGTCAAGGAGCTGGTCAACGAT 4495
GTCCAGCCCGATGGCTTTGTCAGCAAGTTGGTCCTCGACGACGGATCTGCCTCCTCCGCCACCGGAGACATCCACGGCAACATCGACGGAGTCTTCGAGTGGATCTCCCCCGAGGGTGTCCATGTGCGAGTGAGCTACAAGGCTGACGAG 4645
                                                                         GENE III                   Trm
AACGGATACCAGCCCCAGAGTGACCTGCTGCCCACTCCTCCTCCGATCCCAGCTGCCATCCTGAAGGCTATCGCCTACATCGAGGCTAACCCCAGCAAGAACTAAgtgaacccgccgactaggaacatgaaagattggagacagctaggt 4795
                                                       Eco RI
tgagtttggataatttcttaccagttgttttaaatttaaggaaaatgttatcgaattcgaaaataaattaaaccttgcaatataaaccaagtgcatgtttttacaaatctgacagttcgatttaagagaaggctcccggtattatatggta 4945
taagaaggtacaattagaagattaaaagtaatcaaagacactttggccttcattaaattacaattgtgttgttatagtatagtacgaaattaatttaaatacaaaaatctttaaagcatctaaaataaatgtaaacattacaaaaccctt 5095
acctggacaagccgatatctccttgcattaatttcatatttccgaaaactgggttataactagttattattttaagttaagttcataggcagccacaagtaattaaatgttgccaacctgatgcatcccagataagatcgcagtatgatg 5245
aaaacgacgaggaactttttttatatctattatttgtagaggataagggtacacttgaattgttagaacgcatgtcggtattatgggtattaagggtattatgaagcgttttcgaacctaaaaagtatgtatattcatcatcattatcact 5395
agccgagtcgaattttttgtttttatattggtctttttatgaatataactgaaattggcattataagcctagatgtaaaaatcaaattcttcatctttttttttaaccttttttaaatagtcatacacgtaacaaaaaataacacagacttc 5545
                    BglII
cctgaggttacacggttataagatcttgtagtgatttttggagaaatatcaatcaaattgctgtgctttcggattttttgattatattatgatattgtaacttaagtgtttaatagtgattgtataagtaagaatcgtatacattgtatat 5695
gtcaaactcccacgggaatgttcatattgacttaacggaaactagagataaaatatacacacaatgtttttttttattaaacgaaattatttacaataatttaattgactagcaatagtacgctcttcttaggcaacccaatcttatcggt 5845
atcaatttaaactattcttaatatctatgttatttacaaaggttatatgagtaagagtttttgaggaatagaatgtttatgcagattttaatttagtaggatttatgtcaagtcccggtcagtcaagtcttgaggggtggtgaacacagaa 5995
tgttagattccataaacccgttcccagtcatttcgcagatagaaaccaaatgatgctccgaaaggtatgctggatctacaagcggttcgcaaaaaagttttgttttctagttattttttcacctcctaataattaaacttctactatcagc 6145
agcttagacattattcaatcaagttattttttatatgatttgtctggagtaattcaaagttatctgactaaatattccggaagatgttaaattatttcaatgagaaggtggacttacccttttccgagtaacccgattcttttttagaataa 6295
ttacggtagcgatttgcatagacaatagaaatcaaaaagagtgcagcagacgatttttatcgccaccaagcatgtcacttgaaccagtccgtaaaacccaaacgagacctatgctggccgaaatgttaattaaaaacgggttgcatcagct 6445
                                                                                                                   Start
tttgatcagctttaagatttcgtggggagtgcgtataagcgtataaaagccgacgagtgatcccgaattggcatcagtctcacgagttctttagtctgacaatctaaccaagtcaaaATGTTCAAGATCgtaagtatctgaagtttaaag 6595
            INTRON IV                                                    GENE IV
ccggacagttcaatgagtaatcccggaatatcctagCTGCTTGTCTGCGCCCTTGTCGCCCTGGTGGCCGCCAACGAGAATCCCGAGGTCAAGGAACTGGTCAACGATGTCCAPGCCGATGGCTTCGTAAGCAAGTTAGTCCTGGACAAC 6745
GGTTCCGCTGCTTCTGCTACCGGAGATGTCCACGGAAACATCGACGGAGTTTTCGAGTGGGTCTCCCCCGAGGGCGAACACGTCCGTGTGAGCTACAAGGCCGACGAGAACGGATACCAGCCCCAGAGCGACCTCCTGCCCACTCCTCCT 6895
                                                    Trm
CCAATCCCAGAGGCCATCCTGAAGGCCATCGCCTACATCCAGGCCCATCCCAGCAAGGAATAAgcaatcgacacgaccaggacccacattcgaatcggaggtgcaactccaaagaccttgccctctaaccccttagaatttaaacagcatg 7045
cagacattataaatgattatcgagttaggaaataaattcgatactctttggcaacaaatctatttagatatggtcttaatttccctgacggcagcaggagttaccttgtttatggctgatttattttggccgagggaacaaacgctgctt 7195
cagattatcacgaactgtctggatgttacgtgattgatcttagccaatagtaacctgtttaattagcgatacataaagtgaagaccatcaaaccagatttaggtataaattcggtctgtttattacagttttaaatgcaataaaatattt 7345
cattaaacaaaagtcatggctgagcaaaatataaccggattggaattgcttgcgttactcttcatcttcatattgttaaaagaacagtaaagaacggtatagtgaaattttcgaatacttattattattattactcggtttaaatgttgg 7495
tggtacaccgatagaaatttgtaagaaaaaagttaaaataaccatttttttgaaagaatttcggtgccaaaatgagacggtttgagagcgttacactggaaaaaaaacccgatgcaaacatggctttaacgatcgactacctgttataca 7645
ataccccttcacattgtcaatcatctagtataaacttcaaatctaggagtagagagttggtaaaaacatccttgaagatgttaatggactagctgttatcatgattatatcgggcattcaatagtcgagactttatagctctctcttgttt 7795
aatttttttataatttttttctcatgtaaaagttgttaggccacgcttttacagcgttattaataaacaatttgaatgacactttctatttcatccatggtttcagacacttgatgattagttaactaatccagtggtctattttgg 7945
                                                        BglII
caaacatttttttataaacatggaacataatagattggacaaatcgaacaggtgtttctatgcgcctaagatct 8019
```

Table 1. Percent Homology of Cuticle Gene Regions

| Comparison | 5' untranslated | Signal Peptide | in vivo<br>Protein coding | Intron |
|---|---|---|---|---|
| I/II | 95% | 73% (73%) | 91% (95%) | 47% [2] |
| III/IV | 77% | 89% (93%) | 85% (88%) | 59% [1] |
| I/III | 55% [3] | 58% (33%) | 59% (51%) | 33% [2] |
| I/IV | 60% [3] | 67% (40%) | 62% (54%) | 30% [2] |
| II/III | 55% [3] | 51% (20%) | 59% (51%) | 34% [1] |
| II/IV | 60% [3] | 62% (27%) | 59% (54%) | 33% [1] |
| I/ψI | 80% [1] | 76% [1] | 74% [5] | 75% [2] |
| II/ψI | 83% [1] | 62% [1] | 69% [4] | 58% [2] |

Results are indicated in % DNA sequence homology.

The number in [ ] is the number of gaps created to maximize the homology.

The number in ( ) is the protein sequence homology.

# CHAPTER 3

# Two Gene Families Clustered in a Small Region of the *Drosophila* Chromosome

MICHAEL SNYDER and NORMAN DAVIDSON

*Division of Biology, California Institute of Technology*

*Pasadena, CA 91125, U.S.A.*

Running title: Two clustered *Drosophila* gene families

Draft for submission to J. Mol. Biol.

## Summary

Three Drosophila genes which are clustered within 8 kb of DNA at the chromosomal region 44D have been identified and mapped, and the gene cluster entirely sequenced. The three genes are 55-60% homologous in DNA sequence. One gene contains an intron in its 5' proximal protein coding sequence while the other two have none at this position; similarly, another gene has an intron in its 3' proximal protein coding sequence which is not found in the other genes. All three genes are abundantly expressed together in Drosophila first, second, and early third instar larval stages and in adults, but they are not abundantly expressed in either embryonic, late third instar larval, or pupal stages. This gene family lies 11 kb away from another cluster containing four Drosophila larval cuticle protein genes plus a pseudogene. The cuticle genes are all abundantly expressed throughout third instar larval development. Thus, at least seven protein-coding genes and one pseudogene lie within 27 kb of DNA. Moreover, two small gene families can lie adjacent on a chromosome and exhibit different patterns of developmental regulation, even though individual genes within each clustered family are coordinately expressed.

## 1. Introduction

The manner in which genes are organized suggests mechanisms of gene regulation. In eucaryotes, genes that are expressed at the same time of development are often clustered (for examples in *Drosophila*: chorion genes, Spradling *et al.*, 1980, Griffin-Shea *et al.*, 1980, Spradling, 1981; two yolk protein genes, Barnett *et al.*, 1980, Riddell *et al.*, 1981; several heat-shock genes, Corces *et al.*, 1980, Craig & McCarthy, 1980; 68C salivary glue genes, Meyerowitz & Hogness, 1982; histone genes, Goldberg & Hogness, unpublished; cuticle genes, Snyder *et al.*, 1981).

We are studying the larval cuticle genes of *Drosophila* as a model system for understanding the regulation of gene expression in eucaryotes. Five major cuticle proteins and a number of minor species are synthesized and secreted by the epidermal cells of late third instar larvae (Fristrom *et al.*, 1978). Genes for four of the proteins are clustered in 7.9 kb of DNA at region 44D of the second chromosome. A pseudogene also lies within this gene cluster (Snyder *et al.*, 1981, 1982). The four cuticle genes are abundantly expressed in the integument of third instar larvae, but are not abundantly expressed at other developmental stages (Snyder *et al.*, 1981). Thus, these cuticle genes lie in a region of the chromosome which is transcriptionally active at a particular time of development. It is of interest to know (1) how far such an active region extends and (2) if any correlation exists between the pattern of expression of other genes in this region, and the manner in which they are organized.

In this report we investigate the organization and pattern of expression of other structural genes encoded in a 50 kb DNA segment containing the 44D larval cuticle gene clusters. Our results demonstrate that another gene cluster, called the HDL family, lies 11 kb from the cuticle gene cluster. The three genes comprising this family are coordinately expressed during development, but show a pattern of developmental expression completely different from the third instar cuticle genes.

## 2. Materials and Methods

### (a) Enzymes

All enzymes were purchased from New England Biolabs or Boehringer Mannheim, or were prepared by Maria Alonso.

### (b) Lambda and plasmid clones

The recombinant lambda clones $\lambda$DmLCP1 2 and 3 have been previously described (Snyder *et al.*, 1981). Two clones $\lambda$DmLCP10, $\lambda$DmLCP13 were isolated by chromosomal walking by the described procedure (Snyder *et al.*, 1981). The subcloned *Eco*RI DNA fragment pH-11 was previously called pCPB-11 (Snyder *et al.*, 1981).

*Bam*HI DNA fragments of $\lambda$DmLCP13 were subcloned into pBR322 (Bolivar *et al.*, 1977) by the method of Yen and Davidson (1980). The plasmid vector DNA was treated with calf alkaline phosphatase and phenol extracted prior to ligation with $\lambda$DmLCP13 DNA fragments.

### (c) Nucleic acid preparations

Lambda phage and plasmid DNAs were prepared as in Snyder *et al.* (1981) and genomic *Drosophila* DNA as described in Fyrberg *et al.* (1980). Gel isolation of DNA fragments using hydroxyapatite are described in Fyrberg *et al.* (1980).

Undegraded cellular RNA was prepared by homogenizing whole *Drosophila melanogaster* (Canton S) in 4 M guanidinium thiocyanate solution and banding in cesium chloride (Fyrberg *et al.*, 1980). RNAs from larval integuments and from internal viscera were prepared by first squashing second instar larvae with a rolling pin at 4°C. The integuments were separated from the internal viscera by washing the squashed animals through Miracloth using 0°C *Drosophila* Ringers solution. The integuments were collected and directly homogenized in the 4 M guanidinium thiocyanate solution, and solid guanidinium thiocyanate was added to the solution containing the internal viscera to a final concentration of 4 M. The remainder of the procedure is described for whole animals.

The procedure for isolation of poly(A)$^+$ RNA by oligo(dT) cellulose chromatography is described in Snyder et al. (1981). In many cases the RNA was judged to be undegraded by its high efficiency of translation in a rabbit reticulocyte lysate system. In all cases analysis of RNA in gels containing methyl mercury hydroxide (Bailey & Davidson, 1976) showed moderate amounts of the characteristic ribosomal RNA bands.

### (d) Gel electrophoresis and blotting of DNA and RNA

Gel electrophoresis of 0.5 μg lambda clone DNA/lane and 0.3 μg of plasmid DNA/lane was generally performed using 0.7% agarose gels according to Hershey and Davidson (1980). In detailed restriction mapping of subcloned DNA fragments, 1-1.5% agarose gels were used. For preparative, single lane 4 mm thick gels, lambda DNA amounts and plasmid DNA amounts were chosen so as to yield 0.5 μg and 0.3 μg of DNA per 6 mm width, respectively. For gel electrophoresis of genomic DNA, 3 μg of Drosophila DNA was digested with restriction endonucleases and analyzed on a 0.7% agarose gel. In adjacent lanes BamHI digested calf thymus DNA containing haploid genome equivalents of 1, 2 and 3 copies of λDmLCP13 DNA were analyzed.

DNA fragments subjected to electrophoresis were transferred to 0.45 μm nitrocellulose in the case of genomic DNA and to 0.22 μm nitrocellulose in the case of cloned DNA.

RNA was analyzed (1 μg per lane) on 1% agarose gels containing 2.2 M formaldehyde and blots onto 0.45 μm nitrocellulose sheets were prepared according to Thomas (1980). As standards, E. coli rRNA was analyzed in adjacent lanes.

### (e) Restriction maps

Restriction maps were derived according to standard techniques. The following qualifications are noted. (1) Localization of the closely spaced HindIII and XhoI sites that lie to the extreme right of the map in Figure 1 is only to within 500 bp. (2) A closely spaced BamHI site present in λDmLCP13 is believed to lie 100 bp away from the BamHI site at the pL-26, pX-29 junction (Fig. 1). (3) Three ClaI sites, ATCGAT, predicted from the DNA sequence are not observed in digestions under standard conditions. All three overlap the sequence GATC which is methylated in the E. coli host strains used in the study (Backman, 1980).

### (f) DNA labeling and hybridizations

$^{32}$P-labeled cDNA probes to poly(A)$^{+}$ RNA and $^{32}$P-labeled nick translated DNA probes were prepared as described in Snyder et al. (1981).

Hybridization was performed in 50% formamide, 1 M NaCl, 50 mM Tris pH 8.3, 1 mM EDTA, 10X Denhardts, 0.1% SDS plus 10% (w/v) dextran sulfate at 42°C. Probe driven reactions were incubated for 16-20 h, filter driven reactions for 24-28 h each using 1-5 x $10^7$ DPM of probe and approximately 1 ml of solution per 5 $cm^2$ of filter. Filters were washed as previously described (Snyder et al., 1981). Final stringent washes of all blots in 0.3 x SSC at 65°C were carried out (SSC = 0.15 M NaCl, 0.15 M Na citrate).

### (g) DNA sequencing

DNA sequencing was performed according to Maxam and Gilbert (1980). The modifications and observations previously described (Snyder et al., 1982) are relevant. We point out that sequencing across all restriction sites was performed for the entire 9628 nucleotide DNA segment. 80% of the DNA sequence was determined from both strands. In positions of discrepancy, the sequence judged most reliable is presented. In the spacer regions, where sometimes only one strand was sequenced, the frequency of errors is estimated to be less than 1%. Uncertainty in the two nucleotides (NN at positions 1437 and 1438 in Fig. 2) is due to condensed bands; the flanking G and/or C on either side of this sequence may be incorrect.

## 3. Results

### (a) Two gene clusters in the 44D region

The organization and restriction map of 50 kb of Drosophila cloned DNA is shown in Figure 1. Genes I through IV encode four of the five major third instar larval cuticle proteins (Snyder et al., 1981, 1982). ψ gene I is a pseudogene. ψ gene I and genes I through IV are all homologous in DNA sequence (ranging from 59-95% depending upon the particular gene comparison; Snyder et al., 1982). Gene V encodes a nonabundant late third instar larval RNA; its identity is unknown (Snyder et al., 1981). Three clones used in this study are λDmLCP1, 3, and 13 and together contain the entire cloned region.

In experiments described below a region was identified that hybridizes strongly to cDNA probes to poly(A)$^+$ RNA from second instar larvae, early third instar larvae, and adults. This region lies on λDmLCP13 and overlaps with the right end of λDmLCP1. Genes were localized in this region by preparing gel blots containing λDmLCP13 DNA digested with a variety of restriction endonucleases. A representative calf thymus DNA-primed cDNA of poly(A)$^+$ RNA isolated from second instar larvae was used to probe the λDmLCP13 blots. The results are summarized in Figure 1; strong hybridization occurs in two regions which, as shown in more detail below, contain three genes. The three genes separately are called H-44D, D-44D, and L-44D (abbreviated H, D, and L, respectively) and collectively are referred to as the HDL cluster. No hybridization of cDNA was detected to DNA fragments to the left of the BamHI site 3' to H, nor to DNA fragments between the XhoI and EcoRI sites in the H-D spacer, nor to DNA fragments to the right of the SalI site 5' to L. (All combinations of single and double digests of enzymes shown in Figure 1 were tested except BglII digests.) Weak hybridization of cDNA is detected to DNA fragments of the rightmost 5 kb of the cloned region. At least part of this latter region contains sequences repeated in the Drosophila melanogaster genome (see below) and has not been studied further.

H, D and L were further mapped and the orientations of transcription determined by hybridizing $^{32}$P-labeled oligo(dT)-primed cDNA probes prepared from second instar poly(A)$^+$ RNA to λDmLCP13 restriction fragments containing H, D and L (data not shown). The results show that the 700 bp BamHI/BglII fragment harboring the 3' end of H, the 400 bp BglII fragment containing the 3' end of D, and the 800 bp BglII and BglII/BamHI fragments containing the 3' end of L all hybridize strongly to the oligo(dT) primer cDNA. No other DNA fragments of λDmLCP13 were found to hybridize strongly. Comparison of the restriction fragments which hybridize to the calf thymus DNA-primed cDNA with those that hybridize to oligo(dT)-primed cDNA indicated the presence of three genes, encoded on both DNA strands as shown in Figure 1. More precise localization is available using the DNA sequencing results presented in the next section.

### (b) H, D and L are related in sequence and comprise a small multigene family

In order to localize the genes encoded in the D cluster precisely and to understand their structure, the entire region containing H, D and L was subcloned (Fig. 1) and sequenced. The strategy used is presented in Figure 6. The 9628 bp sequence contains 3, and only 3, regions with long open reading frames; these regions correlate well with the DNA that hybridizes to calf thymus DNA primer and oligo(dT) primer cDNA probes to second instar poly(A)$^+$ RNA.

Sequence comparisons of H, D and L show that the three genes are homologous in both DNA (55-60%) and predicted protein sequences (48-53%) (Fig. 2, Table 1). The spacing between the proposed protein coding segments (see below) is 1.7 kb between H and D and 1.1 kb between D and L.

The DNA sequence indicates the presence of a 353 bp intron in H (Fig. 2, position 1322-1323) and a 62 bp intron in L (Fig. 2, position 145-146) for the following reasons. First, these intervening sequences interrupt the DNA sequence homology of these genes (see Fig. 2). Second, the proposed introns interrupt the long open reading frames of H and L; they contain in-frame translation termination codons which would lead to truncated proteins (see below) if the intervening sequences where present in the mRNA. Finally, the ends of the proposed intervening sequences contain splicing donor and acceptor sequences which match well with the Drosophila consensus splicing sequences (Fig. 6). We therefore predict splicing occurs in H and L RNAs thereby generating homologous transcripts of the sizes observed (see below), each with a single long open reading frame. Since we have not mapped the 5' and 3' ends of these genes, additional introns may also be present. Short exons on the 5' and 3' ends of H, D, and L would have to be less than 200 bp to escape detection in the cDNA mapping experiments previously described. Furthermore, the 1515-1572 bp long open reading frames of H, D and L comprise most of the mRNA lengths (1.7 kb) as determined below.

In other examples of gene families it has been possible to recognize the region containing the mRNA start sites by analyzing upstream DNA sequence homology, and by identification of a TATA box sequence (Goldberg & Hogness, unpublished) which lies 20-30 bp upstream of the mRNA start sites (for example, see Snyder et al., 1982). For H, D and L this is not the case; a TATA box sequence is not located at a conserved position in the nearby upstream sequences of H, D and L. For H there is a TATA box sequence 39 bp upstream from the ATG initiation codon in Figure 2. Initiation of transcription 25 residues downstream from this TATA box would result in a 14 bp 5' untranslated region, a size unusual for eucaryotic genes. There are TATA boxes in each of H, D and L beginning 270 to 410 bp upstream of the putative translation initiation ATG codon. If the mRNA start site were to lie 20-30 bp downstream from these TATA sequences, we would expect RNA splicing to occur in the 5' ends of these genes because, in the absence of splicing, the expected transcript sizes would be larger than that observed 1.7 kb, see below); and the presence of multiple (3-7) ATG sequences preceding the one where we predict translation initiation to occur (Fig. 2) is atypical of eucaryotic genes (see Kozak, 1978).

Alternatively, the mRNA start sites of H, D and L may not lie adjacent to a TATA box sequence [the case for yeast cytochrome C (Montgomery *et al.*, 1980) and β globin genes (see Efstratiatis, 1980)]. We note there is a conserved sequence

| | |
|---|---|
| CCACCTTATCGACTaG | H |
| gtgCCTTATCGACTGc | D |
| CCACCTTATCaAtgGG | L |

47 to 72 nucleotides upstream of each gene's ATG codon. The function of this conserved sequence is unknown. The localization of the mRNA start sites for H, D and L must await further mapping data.

At the 3' ends of H, D and L the open reading frame ends in AT-rich DNA sequences, a feature characteristic of 3' untranslated and spacer DNA (see Snyder *et al.*, 1982). For many eucaryotic genes a sequence AATAAA is found approximately 20 residues preceding the poly(A) addition site (Proudfoot & Brownlee, 1976). For H, D and L there are many sequences similar to AATAAA in the region downstream of the proposed translation termination codon.

Immediately preceding the long homologous regions of H, D and L, each gene contains in frame a DNA sequence that codes for an amino acid sequence characteristic of a signal peptide. The peptide is 19-20 amino acid residues long, principally hydrophobic, begins with a methionine and contains a basic residue one or two amino acid residues downstream from the methionine. This structure is characteristic of signal peptide sequences (Kreil, 1982). We therefore predict that these three genes encode similar proteins approximately 500 amino acids long which are probably secreted proteins (see below).

Sequence comparisons of H, D and L with that of the third instar larval cuticle protein and genes [Genes I through IV, major cuticle proteins CP1 through CP5, and minor proteins CP2a, CPX and CP6 (Snyder et al., 1982; Snyder, Silvert, Fristrom & Davidson, unpublished)] failed to reveal any significant protein or DNA sequence homology. As shown below, H, D and L appear not to be expressed in the epidermal cells which synthesize and secrete the cuticle. Their functions remain unknown.

An analysis of protein sequences predicted from H, D and L DNA sequences for their hydrophobic and hydrophilic regions (Kyte & Doolittle, 1982) indicates there is no substantial hydrophobic regions other than the signal peptide sequence. We therefore expect these proteins to be secreted rather than integral membrane proteins.

### (c) Genomic representation H, D and L

Previous studies have shown that the region to the left of the D cluster which is contained on λDmLCP1 and λDmLCP3 (Fig. 1) is present in single copy in the *Drosophila melanogaster* genome (Snyder et al., 1981). We have similarly determined the reiteration frequencies of H, D and L and their surrounding sequences. *Drosophila* genomic DNA was digested with *Bam*HI, and with *Bam*HI plus *Eco*RI and analyzed on an agarose gel. As standards, calf thymus DNA containing either 1, 2 or 3 genomic equivalents of λDmLCP13 was also digested with *Bam*HI and analyzed in adjacent lanes. (One particular experiment also included standards of 0.25 and 0.5 genomic equivalents of λDmLCP13.) Gel blots were prepared and probed with subcloned DNA fragments of λDmLCP13 and λDmLCP1. The results are shown in Figure 3. As seen in the figure, H, D and L sequences are present each at single copy in the haploid genome. On long exposures, weak cross-hybridization (difficult to see in Fig. 3) of the various gene probes with DNA fragments of the sizes expected for the other homologous genes can just be detected. However, to the right of L a 3.7 kb subcloned DNA fragment, pX-29 (Fig. 3) hybridizes to many genomic DNA fragments. We have not mapped the extent of this repetitive DNA further. The maximal extent of this repeated sequence is indicated by the wavy line in Figure 1. It is the only moderately repetitive DNA found in the 50 kb cuticle gene region.

### (d) H, D and L are coordinately expressed, and exhibit a different pattern of developmental expression from the adjacent cuticle genes

The four cuticle genes encoded at 44D are abundantly expressed together during the third instar larval development and are not abundantly expressed at other developmental times (Snyder *et al.*, 1981). A fifth gene, V, which is imprecisely localized and of unknown function (Fig. 1), is nonabundantly expressed in late third instar larvae [2-4 x $10^{-5}$ in poly$(A)^+$ RNA (Snyder *et al.*, 1981)]. We wish to determine the pattern of developmental expression and organization of genes in the entire 50 kb cloned region and, in particular, the manner in which H, D and L are expressed relative to each other and with respect to the neighboring third instar larval cuticle genes. Our conclusions based on experiments described below, show that the only abundantly expressed genes found in the 50 kb DNA segment are the cuticle genes and the HDL cluster genes. Sequences homologous to the region containing gene V and the region containing the repetitive DNA are represented in RNA transcripts, but at a much lower level. Moreover, H, D and L are abundantly expressed in first, second and early third instar larvae and in adults; and in the times tested all three genes are coordinately expressed. They are not abundantly expressed in embryonic or pupal stages tested. The developmental pattern of expression of H, D and L is strikingly different from that of the third instar larval cuticle genes.

To search for abundantly expressed genes in the cloned region we carried out several types of experiments (summarized in Figs. 4 and 5). First, cloned DNAs of λDmLCP1, 3 and 13 (Fig. 4A) were individually applied to nitrocellulose filters and the filters separately hybridized to cDNA probes made to poly(A)$^+$ RNAs from embryos, second instar larvae, two pupal stages, adults, imaginal discs and dissected indirect flight muscle. As positive controls cloned myosin heavy chain (Rozek & Davidson, in preparation), dopa decarboxylase (Hirsh & Davidson, 1981), myosin light chain (Falkenthal & Davidson, unpublished) and cytoplasmic actin (Fyrberg et al., 1980) DNAs were also placed on the filter. (The actin DNA is the only one which hybridizes to all the probes.) As a negative control λ Charon 4 DNA was placed on the filter as well. The results summarized in Fig. 5 show that λDmLCP1 and 13 hybridize strongly to the second instar cDNA probe, λDmLCP3, 1 and 13 weakly hybridize to embryo cDNA and λDmLCP13 weakly hybridizes to 75 h pupal cDNA.

A more detailed analysis of developmental expression profiles of the cloned region in larval and adult stages was performed. λDmLCP1, 3 and 10 DNAs were each digested with restriction endonucleases, separately analyzed on an agarose gel, and gel blots prepared (see Fig. 4A, line 2). Digests were chosen such that each cuticle gene and gene V were on different sized-restriction fragments, and DNA fragments containing H, D plus L, and the repeated DNA are also separated (see line 2, Fig. 4A). These blots were hybridized to cDNA probes to poly(A)$^+$ RNA isolated from second instar larvae, early third instar larvae, late third instar larvae and adults.

The results found (Fig. 4B and 5) were: (1) early and late third instar cDNAs hybridize strongly to each of the cuticle genes I through IV. In addition, weak hybridization (1% relative to the strong hybridization mentioned above) of gene IV to the second instar probe is also found. (2) In contrast, H and D plus L restriction fragments hybridize strongly to second instar and early third instar cDNA probes (10-30% relative to the third instar cuticle genes) and not at all to a late third instar larval probe (Fig. 4B). These same fragments hybridize to the adult cDNA as well. That each gene H, D and L is expressed in second instar larvae and adults was demonstrated by preparing gel blots with the 3' halves of H, D and L separated on different restriction fragments. Figure 4C shows that each of the three genes hybridizes to oligo(dT)-primed cDNAs made to poly(A)$^+$ RNA of second instar larvae and adults. The level of expression of D is judged to be 5-7-fold less than either H or L in second instar larvae. (3) A 5.0 kb DNA fragment containing gene V (Fig. 4B) weakly hybridizes to second instar cDNA. (4) Fragments to the right of λDmLCP10 which lie near or within the repetitive DNA hybridize moderately to late third instar larval cDNA, and weakly to second instar cDNA. These results are summarized in Figure 5.

The expression of genes encoded on λDmLCP10 were further studied and transcript sizes determined by RNA gel blot analysis of poly(A)$^+$ RNA from several stages of Drosophila development. The blots were probed with nick-translated, $^{32}$P-labeled λDmLCP10. As seen in Figure D, prevalent transcripts 1.7 kb in size are found in first instar larvae and much more abundantly in adults. Since, as shown above, H, D and L are the only genes on λDmLCP10 that are abundantly expressed in adult RNA, the 1.7 kb transcripts must be from these genes. Moreover, RNA blots of second instar RNA probed with H, D and L gene-specific probes $^{32}$P-labeled (pH-11, a gel isolated 1.25 kb XbaI/BamHI piece of pDL-31 (fragment D), and a 1.7 kb XbaI/BamHI fragment of pDL-31 (fragment L), respectively [Fig. 4A]) reveals that each of the genes encodes a 1.7 kb DNA transcript (Fig. 4D). A summary of the developmental expression results for the seven abundantly expressed genes in the cloned region is shown in Figure 5.

The 44D Drosophila cuticle genes are abundantly expressed in the integument of third instar larvae (Snyder et al., 1981). In order to determine if H, D and L were expressed in the same tissue during larval development, a crude fractionation of the second instar larval integument and internal viscera was performed, RNA isolated and oligo(dT) primed cDNA probes prepared. Equal amounts of cDNA probe from the integument, viscera, and whole animal probes were hybridized to gel blots containing restriction fragments for the 3' ends of H, D and L as well as a cytoplasmic actin control. The results (Fig. C, lane Int. 1 and 2) show a decrease (~ 2 fold) of cDNA hybridization to each of H, D and L in the integument probe relative to the whole animal probe (lane whole 2nd) when standardized to the cytoplasmic actin control; there is a corresponding increase (less than 2 fold) of hybridization using the internal viscerae probe (compare in Fig. 4C lane Int. 1 with lane viscerae). Therefore, these genes appear not to be expressed in the same tissue as the adjacent third instar larval cuticle genes. Additionally, they are not likely to be involved in *Drosophila* cuticle formation. Although we do not know the functions of H, D and L, we note that the times of expression of these genes correlate well with the times when the animal is feeding.

## 4. Discussion

The results presented above describe a new *Drosophila* gene family. The family is comprised of three genes which are clustered in 8 kb of DNA, are homologous in DNA sequence and, at the developmental stages tested, are coordinately expressed. The sequence data suggest that these genes encode proteins of approximately 500 amino acids; the presence of apparent signal peptide sequences indicate that these proteins are secreted. They are not likely to be integral membrane proteins, because they lack hydrophobic regions except for the signal peptide sequences. The correlation of the time when the genes are expressed (larval and adult stages) with the time the animal is feeding suggests that perhaps these genes are involved in some aspect of the digestion process. Their identity is unknown. It is known that they are not related to the third instar cuticle proteins that have been sequenced thus far (Snyder *et al.*, 1982), and they are not expressed in the tissue expected for *Drosophila* cuticle formation.

The DNA sequence homology of H, D and L indicates that the three genes probably arose by gene duplication events. Their high degree of sequence divergence (greater than 40%) and insertions/deletions within the protein coding sequence suggest that these events were not recent. This is further supported by the lack of intron conservation in these genes, a feature of ancient gene families [*Drosophila* examples include actin (Fyrberg *et al.*, 1981), hsp 70 (Ingolia & Craig, 1982) and collagen (Monson *et al.*, 1982) gene families]. Yet H, D and L are still clustered and have not dispersed to other parts of the genome as have the actin (Tobin *et al.*, 1980; Fyrberg *et al.*, 1980), tubulin (Sanchez *et al.*, 1980) and larval serum protein 1 (Smith *et al.*, 1981) gene families. We also note that H, D and L are expressed together at the same developmental times; and, at the crude level of fractionation discussed above, they are expressed in the same tissue. These facts are consistent with the hypothesis that clustering may be important in the mechanism by which these genes are coordinately expressed. There are many examples of coordinately expressed genes that are clustered [in *Drosophila*, chorion genes (Spradling *et al.*, 1980; Griffin-Shea *et al.*, 1980; Spradling, 1981), two yolk protein genes (Barnett *et al.*, 1980; Riddell *et al.*, 1981), several heat-shock genes (Corces *et al.*, 1980; Craig & McCarthy, 1980), 68C salivary glue genes (Meyerowitz & Hogness, 1982), histone genes (Goldberg & Hogness, unpublished), and cuticle genes (Snyder *et al.*, 1981)]. There are of course also examples of genes which are not clustered, but are coordinately expressed [glue protein genes (see Berendes & Ashburner, 1978); larval serum protein I genes (see Smith *et al.*, 1981) and yolk protein genes (see Barnett *et al.*, 1982)].

Eleven kb away from H, D and L lies another gene cluster encoding four larval cuticle genes that are coordinately expressed in third instar larvae. The H, D and L gene family show a strikingly different pattern of developmental regulation than the four cuticle genes. Thus, if chromatin domains are associated with coordinately transcribed gene clusters, there must be an abrupt transition in the chromatin structure between two independent domains in the 11 kb segment separating the two major gene clusters at 44D.

## Acknowledgements

We are grateful to Dr. Russ Doolittle for the computer analysis of H, D, L sequences for their hydrophobic and hydrophilic regions, and for searching for homology of H, D and L sequences with other protein sequences. We also thank Beverly Bond for generous assistance in the developmental expression studies. We are indebted to Mark Garfinkel and N. D. Hershey for criticisms of the manuscript. This research was supported by Grant GM 20927 to N.D., and training Grant GM 07616 to M.S.

## REFERENCES

Backman, K. (1980). Gene **11**, 169-171.

Bailey, J. M. & Davidson, N. (1976). Anal. Biochem. **70**, 75-85.

Barnett, T., Pachl, C., Gergen, J. P. & Wensink, P. C. (1980). Cell **21**, 729-738.

Berendes, H. D. & Ashburner, M. (1978). In The Genetics and Biology of Drosophila 2B (Ashburner, M. & Wright, T. R., eds.), pp. 453-498, Academic Press, New York.

Bolivar, F., Rodriquez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. B., Crosa, J. H. & Falkow, S. (1977). Gene **2**, 95-113.

Corces, V., Holmgren, R., Freund, R., Morimoto, R. & Meselson, M. (1980). Proc. Nat. Acad. Sci., U.S.A. **77**, 5390-5393.

Craig, E. A. & McCarthy, B. J. (1980). Nucleic Acids Res. **8**, 4441-4457.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980). Cell **21**, 653-668.

Fristrom, J. W., Hill, R. J. & Watt, F. (1978). Biochemistry **19**, 3917-3924.

Fyrberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S. & Davidson, N. (1981). Cell **24**, 107-116.

Fyrberg, E. A., Kindle, K. L., Davidson, N. & Sodja, A. (1980). Cell **19**, 365-378.

Griffin-Shea, R., Thireos, G., Kafatos, F. C., Petri, W. H. & Villa-Komaroff, L. (1980). Cell **19**, 915-922.

Hershey, N. D. & Davidson, N. (1980). Nucleic Acids Res. **21**, 4899-4910.

Ingolia, T. D. & Craig, E. A. (1982). Proc. Nat. Acad. Sci. U.S.A. **79**, 525-529.

Kozak, M. (1978). Cell **15**, 1109-1123.

Kreil, G. (1981). Ann. Rev. Biochem. **50**, 317-348.

Kyte, J. & Doolittle, R. F. (1982). J. Mol. Biol. **157**, 105-132.

Maxam, A. M. and Gilbert, W. (1980). Methods in Enzymology, Vol. 65, pp. 499-560.

Meyerowitz, E. M. & Hogness, D. S. (1982) Cell **28**, 165-176.

Monson, J. M., Natzle, J., Friedman, J. & McCarthy, B. J. (1982). Proc. Nat. Acad. Sci. U.S.A. **79**, 1761-1765.

Montgomery, D. L., Leung, D. W., Smith, M., Shalt, P., Faye, G. & Hall, B. D. (1980). Proc. Nat. Acad. Sci. U.S.A. **77**, 541-545.

Proudfoot, N. J. & Brownlee, G. G. (1976). Nature **263**, 211-214.

Riddell, D. C., Higgins, M. J., McMillan, B. J. & White, B. N. (1981). Nucleic Acids Res. **9**, 1323-1338.

Sanchez, F., Natzle, J. E., Cleveland, D. W., Kirschner, M. W. & McCarthy, B. J. (1980). Cell **22**, 845-854.

Smith, D. F., McClelland, A., White, B. N., Addison, C. F. & Glover, D. M. (1981). Cell **23**, 441-449.

Snyder, M., Hirsh, J. & Davidson, N. (1981). Cell **25**, 167-177.

Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. & Davidson, N. (1982). Cell, 1027-1040.

Spradling, A. C. (1981). Cell **27**, 193-201.

Spradling, A. C., Digan, M. E., Mahowald, A. P., Scott, M. & Craig, E. A. (1980). Cell **19**, 905-914.

Thomas, P. (1980). Proc. Nat. Acad. Sci. U.S.A. **77**, 5201-5205.

Tobin, S. L., Zulauf, E., Sanchez, F., Craig, E. A. & McCarthy, B. J. (1980). Cell **19**, 121-131.

Yen, P. & Davidson, N. (1980). Cell **22**, 137-148.

FIG. 1. The cloned cuticle gene region at 44D. The upper diagram shows the restriction map of the 50 kb of cloned DNA as deduced from the cloned DNA inserts depicted beneath it. Genes I through IV encode third instar larval cuticle proteins CP1 through CP4, respectively. Gene V is a non-abundantly expressed gene of unknown identity. Genes H, D and L are described in the text. The arrow indicates the 5' to 3' direction of transcription. On the right of the map is repetitive DNA; its maximal extent is indicated by the wavy line. To the left of λDmLCP1, only *Hin*dIII, *Sal*I, *Eco*RI and *Kpn*I sites have been mapped. In addition, only the two *Sac*I sites relevant to the text have been indicated.

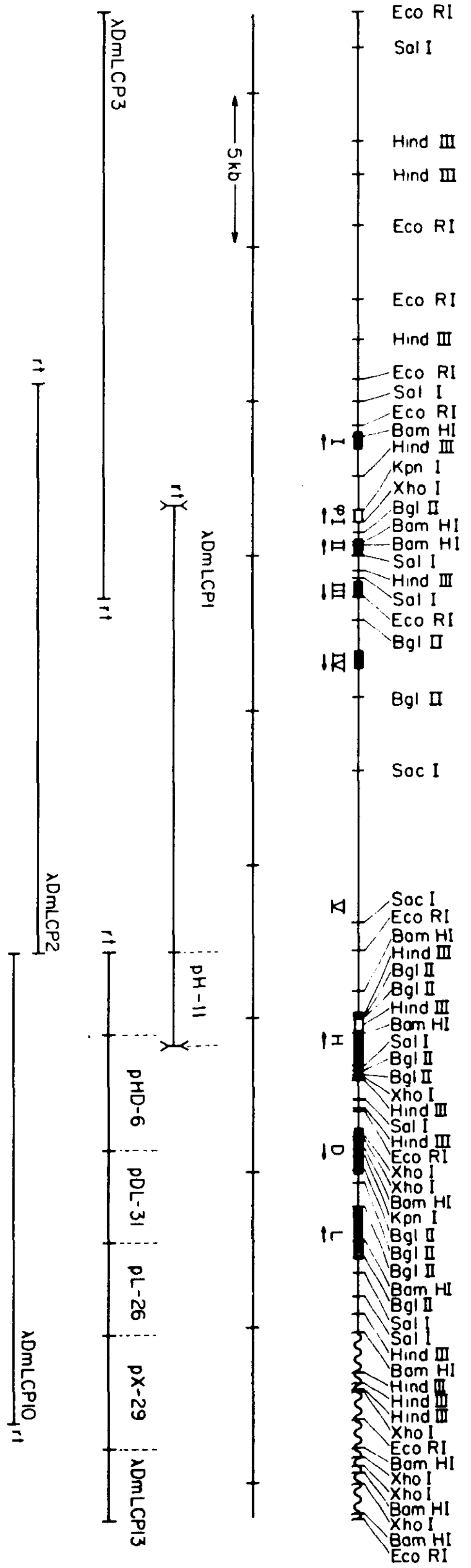

Eco RI
Sal I
Hind III
Hind III
Eco RI
Eco RI
Hind III
Eco RI
Sal I
Eco RI
Bam HI
Hind III
Kpn I
Xho I
Bgl II
Bam HI
Bam HI
Sal I
Hind III
Sal I
Eco RI
Bgl II
Bgl II
Sac I
Sac I
Eco RI
Bam HI
Hind III
Bgl II
Bgl II
Hind III
Bam HI
Sal I
Bgl II
Bgl II
Xho I
Hind III
Sal I
Hind III
Eco RI
Xho I
Xho I
Bam HI
Kpn I
Bgl II
Bgl II
Bgl II
Bam HI
Bgl II
Sal I
Sal I
Hind III
Bam HI
Hind III
Hind III
Hind III
Xho I
Eco RI
Bam HI
Xho I
Xho I
Bam HI
Xho I
Bam HI
Eco RI
I
ψI
II
III
IV
V
H
D
L
5kb
λDmLCP3
rt
rt
λDmLCP1
rt
λDmLCP2
rt
λDmLCP10
rt
rt
pH-11
pHD-6
pDL-31
pL-26
pX-29
λDmLCP13

FIG. 2. Predicted protein coding sequences of H, D and L. The DNA sequences of the three genes are shown according to their homology; identity with H is indicated by a —. Gaps [ ] have been created for better sequence alignment. For the protein sequences of D and L, only differences from the H sequence have been indicated. Two introns indicated by ▼ interrupt H (position 1322-1323) and L (position 145-146). Splicing is required to maintain both the open reading frame and homology of these genes. The 5' and 3' ends of H, D and L have not been mapped; alternative interpretation at these ends is still possible. The coordinate system is based on the DNA sequence of H.

```
   M R P Q S A A C L L L A I V G F V                           G A T E W W E S G N Y Y Q I
H  ATGCGCCCGCAATCAGCTGCGTGTCTATTACTGGCCATCGTTGGCTTCGTG(       )GGAGCCACCGAGTGGTGGGAGAGTGGAAACTACTACCAGATC
   M P K W A H L G   A A L L L I S T T Q E G T A D I D        N A S L
D  ——CAAA-TGGG—CA-CTCG-G——GC-GCTCT-C-TC-AATT-C-ACCACCCAGGAGGGTACT-C—AT-T——T——————AC-CCTCGCTG—————A
   M F K L L V L S C        L A L P S                    L A E V G     K T   Q F
L  ——TT-AA-TTGCTG-TGCTC-CGTGCC-TT———TT-G-C-CTACCTTC-(          )CTG-CG-AGGTT-GC————A——CG——TC-G-T——————
                                                 90

   Y P R S F R D S D G D G I G D L N G V T E K L Q Y L K D I G F T G T W L
H  TACCCGCGCTCCTTCCGGGATTCCGACGGAGACGGCATTGGCGACCTGAACGGGGTCACTGAAAAGCTGCAGTACCTGAAAGACATCGGCTTCACGGGCACATGGCTG
              Q                   K   I   S R   G          E     I    A
D  —T—————————A——CAG——T—C—T————C—G——————A—CA-T——TCG-GAT—GGC—————C—G—A—A—A—————C——C———————
              K            V        I ▼ I   Q Q    P         E     I    A
L  —T—CA-A—T—TAA———AG——T—C————G-G—A—T—C-TA——TA-T—GC—C—————CT—————A—G—A—T——A————A-C——C——————T
       100                             145                                                        200

   S P I F K S P M V D F G Y D I S D F Y Q I H P E Y G T M E D F E R M I A
H  TCGCCCATATTCAAGTCGCCCATGGTCGACTTTGGTTACGACATATCGGACTTCTACCAGATCCATCCCGAATATGGAACCATGGAGGACTTTGAGCGAATGATCGCC
          T        S              N     D    D    I F     L       D D L   V
D  ——————T——T-CC——————TCG—T—C—C——————C—TA————T—TG-C——G————GATT-TC—C—TC-C—————T—C—TGACC-T—A-TG
          T        A          V A   L K G    D    I F                A L L
L  —C—G—C————CC—A——————CT—T—C—C——————TG-GG-T—TC-TA-GGGT—TG—————AT—TC—C—A————————T—C——GCCC—C-T—A
                                250                                           300

   K A K E V G I K I I L D F V P N H S S T E N E W F T K S V D S D P V Y K
H  AAGGCCAAGGAGGTGGGCATTAAAATCATCCTGGACTTCGTACCAAACCACTCAAGTACCGAAAACGAATGGTTCACCAAGTCTGTGGACAGTGACCCCGTCTACAAA
   E     S L  V                            D      V       E         N R E D G    D
D  G—————TC-C——G——G—————T—C—————T————C————————G—GA—G——TG—————TGAA——————C——A-TC-C—AGAT-G————G-C
   R       L D                           T     D    C D      I R    A A G E E E
L  CGC—————A——C——A——C——————T——————T—G—C————A-C——GA——GTG——C——————TACGA—G-CA-CTG-C—GGAG-AG—————G
                                350                                  400

   D F Y I W H D G K I N N E T G E R E P P S N W N S E F R Y S A W E W N E
H  GACTTCTACATCTGGCACGATGGCAAGATCAACAACGAGACCGGTGAGCGGGAGCCGCCGAGCAACTGGAACTCTGAGTTTCGCTACAGCGCCTGGGAGTGGAACGAG
        V   D        L  E      A   D              V    V   S G P M    T
D  ————————G-G—G—————A—AC-G—TG-G—A—T————CAA-A—T—T—ATCT—T——GT-AG—T——CA—GGACC-ATG——AC———————————
        V    P T       V V          I   Q        T        V    V      G    M    T
L  —T————TG-T——CAACC—————G-GGT———(      )—AATC—CC————A—C-C————————GT——C-T——C—GGATC-ATG——ACC———————
                              450                                   500

   V R Q Q Y Y L H Q F A I Q Q A D L N Y R N P A V V N E M K N V I R F W L
H  GTGCGTCAGCAGTACTACCTCCACCAGTTCGCCATCCAGCAGGCCGACCTCAACTATCGCAATCCGGCCGTGGTTAACGAGATGAAGAACGTTATTCGCTTCTGGCTG
   K          F          Q V K    P        F T      M    R E H   L D   L K
D  AA—C——————————T——G——————TCAGG-GA————C-G—T—A————T-AC——————AATG—CAGGG-GC-C——CT-G-T—CT-GAAG—T——————
   E    A                H A K    P                     K     E A       Q    L
L  -AA—G—GC——————————————A——CA-GC-A-A——C————T—A—————C——————C—CAAG—T—G-A-CC—————G-T—AC-G——————————
                       550                                     600

   G K G V S G F R I D A V P Y L F E V D L D R Y N Q Y P D E P L T N D S V
H  GGAAAGGGAGTATCCGGATTCCGTATAGATGCAGTTCCCTACTTGTTTGAGGTGGATCTCGATCGCTACAACCAGTATCCGGACGAGCCTCTGACCAACGACTCCGTG
   D R    D                        H I Y     H R N A D G S                        V   G
D  -ACCGC————TGAT—————T——————C——————————AC—A-C-AC——(  )C-C-G-A—GCAG—GGATCC——————————T—A—G(       )-TTAG—GT
   R     A Y                      H V Y     I P A    A D G N W                       R N E
L  C——————G-CT-A—C———C—C—C—C————TC-TG-C-AC—AA-TCCAGCT——GC-G—GG-A-C-GG—C—T—A—G(        )CGGAA—A-
               650                                    700

   N C P D P D D H C Y T Q H I Y T Q D M P E T I D M V Y Q W R E L V D E F
H  AACTGTCCCGACCCCGATGACCACTGCTACACGCAGCACATCTATACACAGGACATGCCGGAAACAATTGACATGGTGTACCAGTGGCGTGAGTTGGTGGATGAGTTT
   W G S     N A Y D   H D              K    Q    A    V   L M    E                F L    N Y
D  TGGG-CAGT——————A-C-C-T—GA———CATG-T——————————C—CA—————CA—————CC—GG-G—TC-CA————G-A——————————TT————CA-C-A-
   A V S      E     Y T   L               T      Q          L E L       A F   D V I E   I
L  GCGGTGAG———T————G——T-TACT——CT————————————C—CACC—TCAA—C—G—CT-G—GC————T——GCT-TC—C—TG—A-C—G—AA-A
          750                                       800

   H V E N G G D K R L L M T E A Y T S F E N I M T Y Y G N G V R N G S H I
H  CATGTGGAGAACGGCGGAGATAAGAGGCTGCTGATGACGGAGGCGTACACCAGCTTTGAAAACATCATGACTTACTATGGAAATGGAGTCAGGAATGGCTCCCACATT
   R A Q          S    V    L A         S    V    T L S A    F      S T H Q    T Q L
D  -GG-C-C————T—A—C——TCAC—G-T—CC—G-C——————C—TT-TTC-G-G—G-CAC-A-GTG——————TC—————CA-CAC-CATC-G—TA————GC-G
   D A   L         D    V   L          S P L    V L   Q                      T H L      Q
L  G-C-CC——CTG—T—C—CG-CC-CG-C——C-C—C———————C——T-TCCTC-G——GTTT-G——CAG———————————————CACTCACCTG——————G—C
   850                                      900

   P F N F D F L T S I N N A S K A G E Y V K H I K K W M D A S P E G V Y A
H  CCCTTCAACTTTGACTTCCTTACGAGCATCAACAACGCCTCTAAGGCCGGAGAGTACGTAAAACATATTAAGAAGTGGATGGACGCCAGTCCCGAGGGCGTTTATGCC
    M     Q L M Y L S G Y S T A K D V V G S I D Y W M N T M W K    H Q T
D  ——A-G—————CC-G—GA-GTATCT——GTGGAT——T——A-CGCCAAG-AT-TTGTG-G-TCTATCGACT-CTG-AT—ATAC-ATGT-GAAG————CA-CAGACG———
            E L   A Q   S Y S   D    Y H    S E L    H N    L    N M       Q V
L  ——————————T—C—GC-G—GG-CCAG————G-T—AG——CG-T——TACC-T——TCGG-G-TG——CC-C—C——C————AA—TG—A——A—TCAGGTG—T
                                       1000                                          1050

   N W V L G N H D N K R V A S R F G V Q R T D L I N I L L Q T L P G H A V
H  AATTGGGTGCTGGGCAACCACGACAACAAGCGAGTAGCCTCCCGATTTGGTGTCCAGCGCACCGACCTTATCAACATCCTGCTGCAGACTCTGCCCGGCCACGCGGTC
                T N      D   M   A H K V   L   V I V N A            A S
D  ————————TG-T—T—T—————T-C——C—————G——GAT—TA-G—A-CT—CAAAGTG—T—AC-A—TG-GA-TG-AA-CG-C—————A—TGCTT—————A
     F          Q S  I G      L   A D    I   A C    M I I L G               V S
L  —C——————T-C——————————C-G-GT—CA-T-G—————————G—C-C-G-T————T————GCCTG——————GA-CA-A-T-GGC—————————————GTGAGT—G
                             1100                                 1150

   T Y N G E E L G M T D V W I S W E D T V D P N A C N S D P D N Y Y A R S
H  ACCTACAATGGAGAGGAGCTGGGCATGACCGACGTCTGGATCAGCTGGGAGGACACCGTTGATCCGAATGCGTGCAACTCCGATCCCGACAACTACTACGCCCGATCC
      Y     I     S N  D V E C T G D S C E D
D  ——————T-C—C——————AA-A——————T—A————GGACG-GGAG—CACC-G-GA-TC-TGCGA-G—(                                        )
      Q       M                          Q    Q   N E Q E F E R L T
L  ——————C-G—C———————A————T——————T——————————————————————G—C—TC-A—C——C-A—GA-CGAAC-GG-G-TTG-GCG——TCA-T
                    1200                                      1250

                                  ▼
   R D P A R S P Y Q W D A S S K A G F T S A D H T W L P V A D D Y K T N N
H  AGGGATCCTGCCCGATCTCCCTACCAGTGGGATGCCTCCTCCAAAGCAGGCTTCACCAGTGCCGATCACACCTGGTTACCAGTGGCTGATGACTACAAAACCAACAAC
     G E  T   M     T  G K N    D    S D G E S                 L S P E   Q R Y
D  C-C——GGG-AG—CA-G——ATG——————ACG—TGGAAAG—T—C-A————T-TGA—GG—GTC——————————C-G—TC-AAG-CCC—A——TC—CGGT————T
        V   T    F       S D E V N       S N   S V              S N       L V
L  C-C—C—A-TG—CA———A-T————————AGC-A-GAGGT——C—C—G———T——AC———TCGGTA———————C-G—T——————CAGCA—————————GTTGGT———T
                1300                     1322                     1350

   A L Q Q L R A P R S H L Q I F K K L 7 R V R K E P S F R Q              G E L
H  GCCCTGCAGCAGCTGCGAGCTCCACGTTCCCACCTGCAGATCTTCAAGAAACTGGNNCGCGTGCGCAAGGAGCCATCTTTTCGCCAG(              )GGAGAACTC
   V Q T E R G V S      S    N        G    Q E L K S S S A    L A F K E D        G F
D  -TG-A-ACTG-A-GCG-T-T-T-CA-A—TTC————A-C—————T——GGT—TCAAGAAC-TAAG-GCAGTT-CG-C—————TTGC-TTCAAAGAGGAT————GTT—
   V K K E R G I A L          N V Y    Q    R A L     D       T L K                         D V
L  -TGAA-A—G——GTG——ATAG-C-T-AG——————A-CG——A————C-G——CGGGCTC————TG-T——————TA————GAAG———(               )——T——TG——
         1400                                         1450

   N I Q A I D D D V I I Y S R Q V V Y Q S L R S L S F D Y Trm
H  AACATTCAGGCAATCGACGACGATGTGATCATCTATTCACGGCAAGTCTATCAATCTTTAAGATCTTTAAGCTTCGATTACTAA
   S Y E    V T E Q    L Q I I    Y    K    I     F Trm
D  TC-TACG————GG-GACG——AC-G—TC-ACAGATCATT————T-C——GA-G————ATCC-CTTC-AA
   S V T      G P N    L A F K     Trm
L  TC-G-AACA—C—T-GTCCAA—————C-GGC—TCAAG——T—
    1500                                 1550
```

FIG. 3. Genomic representation of H, D and L and their flanking sequences. Genomic Drosophila DNA (3 µg) was digested with BamHI and BamHI plus EcoRI and fractionated on a 0.7% agarose gel. As standards, calf thymus DNA plus haploid genomic equivalents of 1, 2, and 3 copies of λDmLCP13 DNA was digested with BamHI and included in adjacent lanes. For the pH-11 experiment, additional standards of 0.25 and 0.5 copies were included. Gel blots were prepared and probed with the subcloned probes indicated. Units are given in kb.

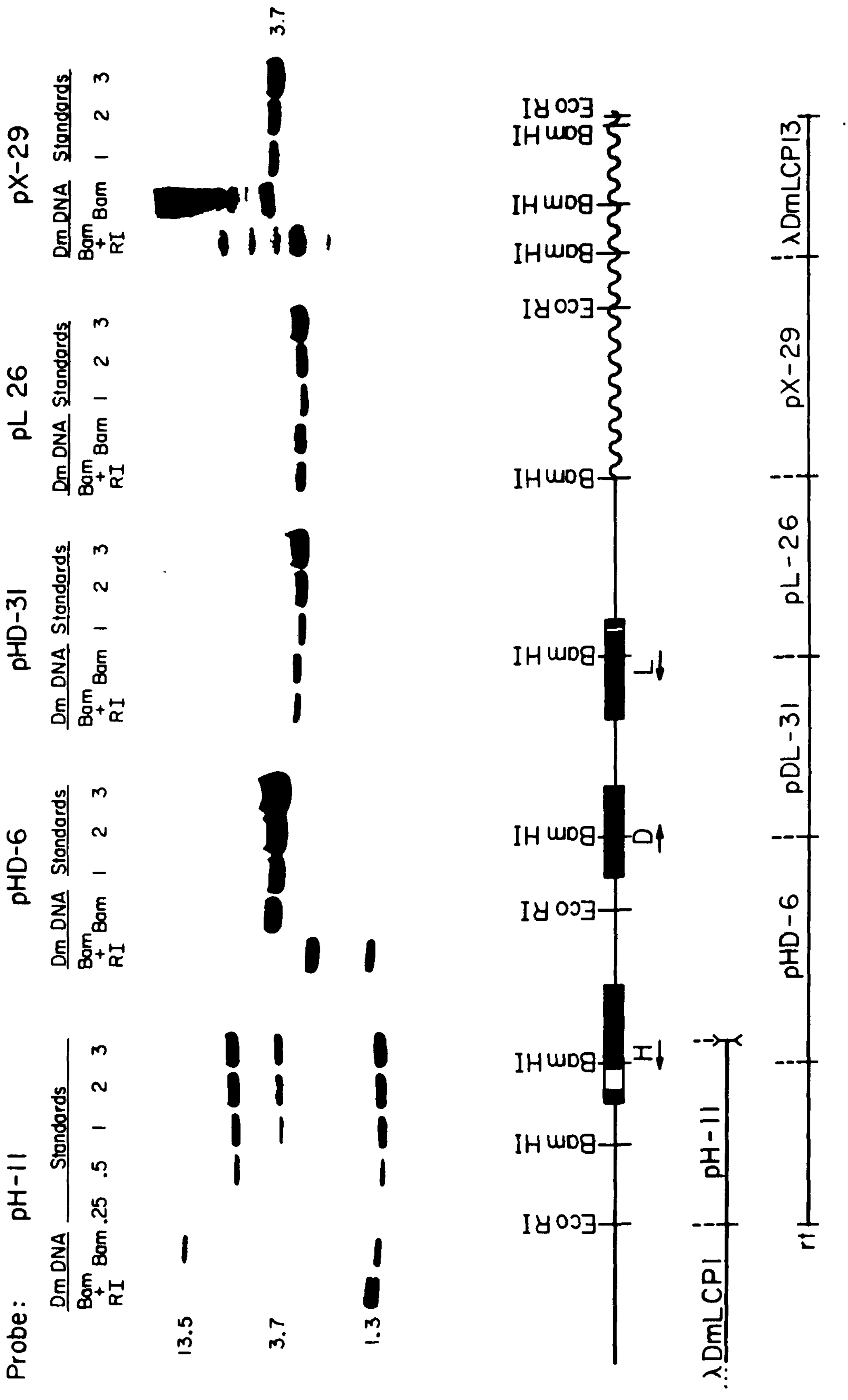

Probe:
pH-11
Dm DNA
Bam + RI
Bam
Standards
.25
.5
1
2
3
13.5
3.7
1.3
pHD-6
pHD-31
pL 26
pX-29
3.7
Eco RI
Bam HI
Bam HI
H
Eco RI
Bam HI
D
Bam HI
L
Bam HI
Eco RI
Bam HI
Bam HI
Bam HI
Eco RI
λDmLCP1
pH-11
rt
pHD-6
pDL-31
pL-26
pX-29
λDmLCP13

FIG. 4. Developmental expression of cloned genes at region 44D. (A) Clones and probes used for analyzing the developmental expression of genes encoded in the 44D cuticle gene region as described in text, and in subsequent panels. (B) λDmLCP3 was digested with HindIII plus EcoRI, λDmLCP1 with HindIII, EcoRI plus SacI and λDmLCP10 with XhoI plus SalI (see panel A for the DNA fragments generated). The resulting DNA fragments of each clone were separately fractionated on a .7% agarose gel, gel blots prepared, and the DNA blots of each clone were collectively hybridized to cDNA probes of 2nd instar RNA (50-54 h after egg laying; left panel), early third instar RNA (72-75 h after egg laying; right panel), late third instar RNA and adult RNA (not shown). The genes contained on the hybridizing fragments are indicated. (C) Gel blots of H, D and L specific fragments probed with cDNA probes. 10 μg of pDL-31 was digested with BamHI + XbaI, and 10 μg of pH-11 plus 10 μg of a subcloned cytoplasmic actin DNA (Fyrberg et al., 1980) was digested with EcoRI. The resulting DNAs were combined, and fragments containing the 3' ends of H, D and L were separated on a 0.7% agarose gel. Gel blots were prepared and probed with oligo(dT) primed cDNA to poly(A)$^+$ RNA from the stages and tissues indicated which from left to right are: Int. 1 and Int. 2 = two second instar integument preparations, whole second instar animals, the internal viscera, a second instar larvae and whole adults, 0-5 days after ecolsion. The top portion of the figure containing actin is a shorter exposure (3-fold) than the bottom portion. (D) Developmental RNA blots using λDmLCP10 as a probe. Left panel. 0.5 μg each of total poly(A)$^+$ RNA from Drosophila at various developmental times was isolated and analyzed on a 1% agarose gel. From left to right lanes are: Embryo 0-4 h, Embryo 13-17 h, First instar larvae 40-44 h, Prepupae 0-5 h, Pupae 75-76 h, and 0-5 day adults. Right panel. RNA blots of 0.5 μg of second instar poly(A)$^+$ RNA probed with H, D and L specific probes, pH-11, fragment D, and fragment L (see A). Units on the right indicate the positions where 16S and 23S E. coli rRNA migrate, respectively.

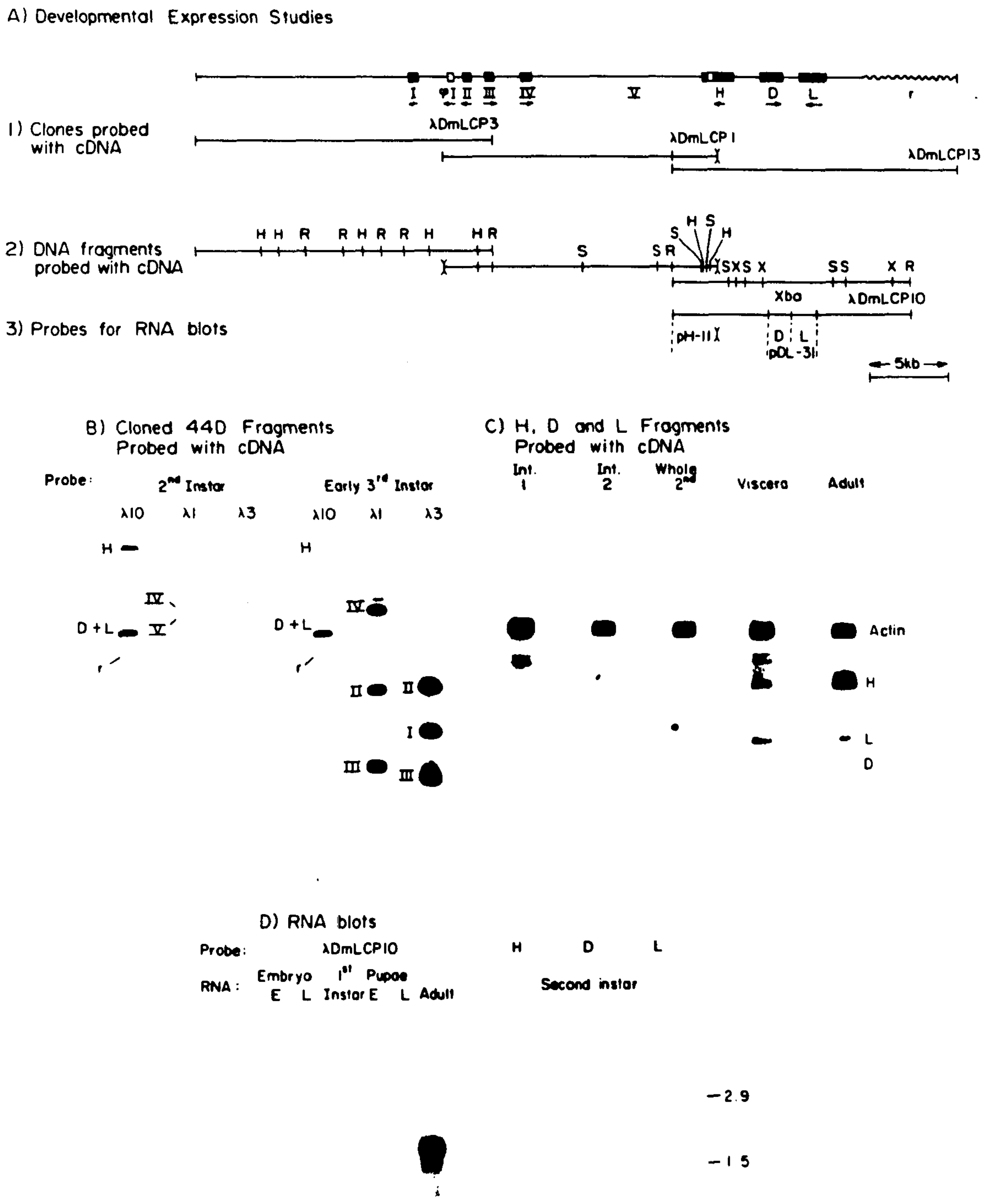

A) Developmental Expression Studies
I
ΨI
II
III
IV
V
H
D
L
r
1) Clones probed with cDNA
λDmLCP3
λDmLCP1
λDmLCP13
2) DNA fragments probed with cDNA
H H R R H R R H H R
S S R
S H S H
S X S X
S S X R
Xba
λDmLCP10
3) Probes for RNA blots
pH-11
D L
pDL-31
5kb
B) Cloned 44D Fragments Probed with cDNA
Probe:
2nd Instar
Early 3rd Instar
λ10 λ1 λ3 λ10 λ1 λ3
H
IV
D + L
V
r
H
IV
D + L
r
II II
I
III III
C) H, D and L Fragments Probed with cDNA
Int. 1
Int. 2
Whole 2nd
Viscera
Adult
Actin
H
L
D
D) RNA blots
Probe:
λDmLCP10
H D L
RNA:
Embryo E L
1st Instar
Pupae E L
Adult
Second instar
–2.9
–1.5

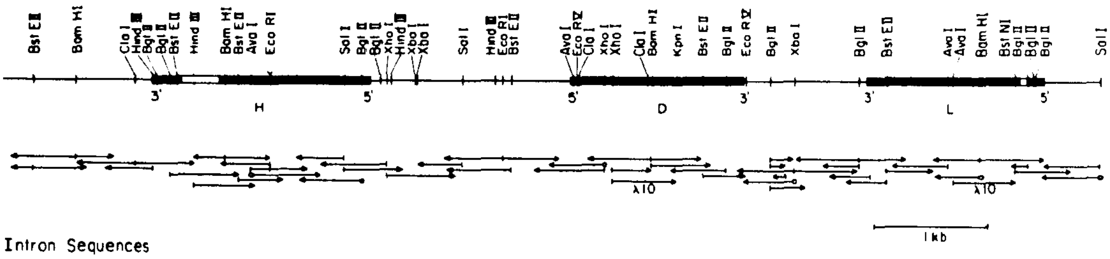
FIG. 5. Summary of the developmental expression results for genes in the 44D cloned region. From the data shown in Figure 4 the developmental expression of genes and areas indicated is deduced. The scale shown is from ± (barely visible over background) to +++ (0.3% of total poly$(A)^+$ RNA for cuticle genes, Snyder *et al.*, 1982). The boxed areas indicate positive results obtained without using H, D and L gene specific probes; any or all of the genes within the box may be expressed at the time indicated.

## Summary of Developmental Expression Results

| | Embryo, 0-4 h | Embryo, 13-17 h | Embryo, 16-20 h | First instar, 40-44 h | Second instar, 50-54 h | Early third instar, 75 h | Mid third instar, 90 h | Late third instar | Prepupa, 5-6 h | Pupa, 33 h | Pupa, 75-76 h | Adult, 0-5 days | Indirect flight muscle | Imaginal disc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | - | - | | ++ | ++ | +++ | - | - | - | | - | +++ | - | - |
| D | - | - | | | ++ | +++ | - | - | - | | - | +++ | - | - |
| L | - | - | | | ++ | | - | - | - | | - | +++ | - | - |
| Gene I | - | | - | | - | +++ | | +++ | - | - | - | - | - | - |
| Gene II | - | | - | | - | +++ | | +++ | - | - | - | - | - | - |
| Gene III | - | | - | | - | +++ | | +++ | - | - | - | - | - | - |
| Gene IV | - | | - | | ± | +++ | | +++ | - | - | - | - | - | - |
| Gene V | | | | | ± | | | ± | | | | | | |
| Repeated DNA Region | | | | | + | | | ++ | | | | | | |

FIG. 6. Sequencing strategy, intron and spacer DNA sequences. (A) Sequencing strategy. A restriction map of the region containing H, D and L is shown. The dark box indicates the protein coding sequence and the open box the introns. The DNA sequencing strategy is such that a vertical line ( | ) or circle (O) indicates $^{32}$P-labeling at either a 3' or 5' terminus, respectively. The arrows indicate the direction and extent of sequencing from the labeled terminus. (B) Predicted intron sequences of H and of L. The intron/exon boundaries are compared with a *Drosophila* consensus sequence (Snyder *et al.*, 1982). (C) Spacer sequence plus the remaining sequences not present found in B or Figure 2. The 5' to 3' sequence is presented minus that previously shown. Start and Trm indicate the DNA coding for the putative translation initiation and termination codons shown in Figure 2.

A) Sequencing Strategy



B) Intron Sequences

```
H Intron

 GGTPAGTP      Dm consensus                                                                                                       117
CAGGTATGAACTTATCTATATTTACTCTGCTTAGTATGAACAATCTACACTGAACTAAAACATTTGGTTAAGTGCGTTACAAAGACTCCGAAATCTCACCTCACTGAAGGGCCATCTGAC
                                                                                                                                  237
ATTCTATTGGAATACAAATGCTAACATTTTTAAAAGACATGACGAATGATAGCCGTTTGAATATAATAGCTAATCAAATACATTTTATATATTGACGTTGCAGTGTGATAAGAAGCTTTA
                                                                                                                               353
TATGTTCTGATAAGAAAAACATTTGTGATACTTCTATTGTACTAATCTAAGTTTTACTTATGTAGCACTGACATTGCGGTTCTTATTGATATTCCATTTATATATACCACTTTCAGGCT
                                                                                                              Dm consensus   tcYaYYYAG

D Intron
                           GGTPAGTP   Dm consensus                            Dm consensus   tcYaYYYAG
                          TAGGTGAGTAGGCGGGTCGAAAGCGTACTACTTAGCCTATTATCTTACTTCACTCTTGCACTTAGGTA
                             1                                                               62
```

C) The HDL cluster sequence

```
5'
GAAAATGGCATTTTTGGACACTCTGCTGTCATCCAAAGTGGATGGACGTCCTCTGACTTCCCAAGAGCTTAACGAGGAAGTTTCCACCTTTATGTTTGAAGGGCATGACACTACTACATCTGGTGTAGGTTTTGCAGTCTACTT   144
ACTTTCTCGACATCCAGATGAACAGGTAACCGATTTGAAGAACTAGTTTTGACGAAGCGAATTTAATAATATATTACTTTCTGTTTGCAGGAAAAATTATTTAACGAGCAGTGCGATGTGATGGGCGCTTCTGGACTTGGTCGA   288
GATGCCACGTTTCAGGAGATATCCACAATGAAACATTTGGATTTGTTTATAAAGGAGGCGCAACGTCTTTATCCGAGTGTCCCTTTCATTGGTCGCTTTACTGAGAAGGACTACGTAATTGGTGAATCTTCGAAACAAATTTAG   432
GCGTTATCAGTAATCAGGATTATGTTTTTAGATGGTGACATTGTGCCGAAGGGGACTACCTTGAACTTGGGCCTTCTTATGTTGGGATACAATGACCGTGTTTTTATGGATCCTCACAAATTCCAGCCCGAACGCTTCGACCGT   576
GAGAAACCAGGGCCGTTTGAGTACGTGCCCTTTAGTGCTGGTCCCCGAAATTGCATTGGTCAGAAGTTTGCACTGCTGGAGATCAAGACTGTGGTGTCCAAAATCATTCGAAATTTCGAGGTGCTGCCAGCTCTGGATGAGCTC   720
GTTTCCAAGGATGGCTATATAAGCACAACTCTAGGTCTTCAGCCGGCAGAGAAGAAGAGCCGTGATGCTCACAACCATAAGTATGATCCAATTTTGTCGGCATCCATGACTCTGAAGTCCGAAAATGGTCTACATCTACGCATG   864
AAGCAGAGGCTTGTATGCGATAGTACATGAATAAACACTTGCATGCTTATGGGCTACGACTTTGCAAACTTTGATAAGTACATGTATAATTCACATTTTATTTATCAGCGCTACTTTATTGTCACTTATTTGAATAAAATTAGC   1008
TTAGACAGCCACTAGGACGGTGCCCACATAGGGATTGGCAACGAACTCCGTAGATTTGATAACATCGCCATCGATGTACTGGGAGCTCAAGGACGTGGTAATGACCTCCGCCTGGGTGCCCAGTTCGTAATACTTGGTCAGGTC   1152
                                                                                      Trm
CAAGGTCTTGGAGGTGCTACCCAGGTTGAGAACGATTACATAGAGATCGCTTCCGGTCTTTTGTCTATTCGAGAGACAAACGTTTA

                                                        H Protein Coding Sequence···
Start
CATTTTGATCGATTTTGTTCGGGCTGGAGAGCTGACAGTGCCTTTTATACTAGTCGATAAGGTGGCTCCAATCGAGATCTGGCCATTTTGAGTGGCCATTTGAATTGCTGTGCAAACAGCTCCCATGAGTTCTCAGGGTGCTCG   3298
AGTCGAACCAAAACTGTCATTGTTTTCGTACTCCCAGCGAAGCTTTACGTAAAGCGATTAGCCCACTCCCGAATGTCTTCTGATTGCTGGAAAAATACCGATTTTTGCGACAAATATCATCCGGAATTATTTGTATACCCGTTA   3442
CCGGTAAGTGTAAGCAGGTATACAGGTTCGGCGAAAATAAGGTATAATATCGGAATTATGGGAAATAAGAGAGCCAAACAGTTGAGAATACGGCTTTTGCTTCTAGAAAAGCAAACACATCTAGAGATAGTTCCAAAAGATTGC   3586
CACACGCTATACATTTAAAGTAAATATGAAGTGAATTAAATTTTACATTCTGCTGCTGGGTTCCACTCTGAGACCGAAGGTTTTCTGTTTGCAATTTGAGTAATGGGTAACATATAGCGCAGCACCTGTCTTTTCTGTTCCTCC   3730
TTGTTTTTGTTATGGTAGGCCTTTCAAGTGGACATTAGTCACCTGATAATAGCGCAGACAAATCCTACTCCTTCAACTACAAGCAAACAAGCAACAAGCTGGAAAACCATGATTAATTAACACTTAACAATGCTGTGGAAATGT   3874
TTAAAAGATAACCTTGTGGAAGTTTTGGGGGCATTTTAATCGCGTGACAATACATAATCATCGTTGCAATACGGCCTAAACTATTGTCAGTATCAGCGTAGGTCGACATTCATATGATAAGCGTTTTCGAGGAAAATATTTCCT   4018
GCGTTTATCTAACTCTATAAAAATGTCATATTTTACTCGCAGTAACTTGTACAGTACAGTAACTTGTAAAGTCCACTTAAGATTGTCCGTCCGCCTGTCTGTCAATTACAATAAATGCATCCTTCATCAGGTTGCGTTAGTGTA   4162
GTAAACTACCTTAAAAGGTAACTCAAAATATACATGTATTTGTTAGCTATATTGATTGTATGAAACCACAGATGGTTTAGATAGTACCCGTACCTCAAAAAGCTTATCATGAAAAATACAAAAACATCCTTTGAATTGTGGCTA   4306
GAGTGGTATCAAGAATTCTACATGCAATTTATGATTTCATCTTTTGTGAACTCTGGAGCATATTGAGGTGGAATTGTAAAGGTAAAGCCTAAATGGTGACCTAATTAGTATTATAAATCTAAACATTTTTTGTTGTCGAGCAAA   4462
AAGCATATGGTATTTCAAAAGTATTTTGGCCGTGCACGCATTATTGCCCATTTATTATCTGTCTTGCTCAAATACACTGAAATATGTTTGTAGTAATTTTTATAAATTGTAAAATATATTATATTCTATAAATTCTATATTATA   4606
TTTGTAATTAAACCCGTTACCCTACGAATATAAATACTAAAAGATATTGATGCAGCTTGTGATTCTCAGGCCAGTCATCAATTTATAACTTTAACAAACGTACTTTATTGCACTTTTACGACATTTCGCGGTCTAATGGAATGC   4750
CCAACAAAAAAGAGAGCATTCGAGAGCGGGAAAGACCAGTCCACTGTTATTTAATCTAGCAAATATGCGATAAAACTGATAGATTACCAGTTGGCCGGTGCAATTGTGCCTTATCGACTGCTCAATCGCTACATGGCAATTGGT   4894
                                        Start
ATTAGCAAATTGAGTTTGGCCCACTCAAACGCTATCGAAATG

                                                        D Protein Coding Sequence···
Trm
TAATCCTCTAGCTGATCCTAATATCCTGATATCCTACAGCACAAACAAGATCAGCGAGGAGTATCGAATCCTGGTCAACATGGGCAACGGTATGGAAATCCTCGATGGGCTCGCCACCAAAACCTACGAGTATGTGCTGGCCAC   6589
TGCCTACTCCACACACTATTCGGGGTAGGTTTTCTCTACTTTCTTGATTTCTATACTTTTAACTAACTTAAAACGATTTACAGGCAAAAAGCAGATCTGTCACAGAGAATCATTCTCATGCCCTACGAAGCAGTCGTTCTACGC   6733
TGGTTGGCTTAACTTCTTGTACTTATTTTGTTGTATCAAGAATTACTTCTTTAGTTCGTGTTAATCAGTACGAATGCCACTGGGGTCCCTAATAAAGATAGGCCCTGTTTCGACATATGTTTTCCCAGAAATATATTAAACCGA   6877
ACTTTGAATGCCAATTACCACATCTAGATTTTATAGACACCCACTTAAGATATAGGCATTTGGCCTGAGCTGTCGCTTGGTTATCGCTATGGATTAATCTGCGAATAAACGTGATCGATAGTAATCGTACGAATTTATTTATTG   7021
CCTAAATGGCCAATAAATTAAATGGCTATTATTCAAAACTTCCAGTGGGTCTGTTTCTAATCAGTGCTTATACAAGTTGGCATGGGGCCATCGTATAGATAAGCACAGTTAGTTAAGTAATCCAACCCTATTAAAGAAGTTATG   7165
CGATCAAAATATTTTTGTGTTATTCTTTATTAAGGATCTAATATATTAGTTTAAACCGTGCTCAGCACAACTGCTTCCTTGGGCAACAGCAGCACGGAGTTGGCAAAGGTGAGGTCGCTAAAAGGATAAAATAGTTAGTGTTGT   7309
ATGTTTAAAGATATTTTATGGTGTCTAGTGCTTACTTCTTGCGACGCACACTCTTATCATTCACCACCACATATTGCAGTTGCGTGGTTATGGAGGTGAAGACAGAGTCCAAGTTGATCGATTCCACATCATCGTTGATGTTGA   7453
                                                                                              Trm
TGAGGGTGATGTACGACTTGTAGCCCGCCAAAGATCTGTTGACAAGGAAAACCGGTCAATATCCAGTCTTTGTGTAAGTTCTATGATTGTTCCAGCTTA

                                                        L Protein Coding Sequence···
Start
CATGGCTCCCACTAAATGGATCGATTGCCAATGGCCAGAGGGATTAAAAGGTGCCCCTATTTAAGCATACCGCTCCCATTGATAAGGTGGTGATAAACCCACCATATAGAAAACTATTTCTGATTACCTCCGTGGAGCTGGCAG   9270
ATTTACGAACCCACGACACTACGAATGGGTTCTCATGATACGCAGGCTGTGGAATTTATTGGAGTCAATTTCATGGTTGTTTTAGCCAGCTCACGCCTAGAGATCATGAGACAGATTGGCTGGTAAATGAATCCAACTGCATAG   9414
CCATATCTTATCGGTGGACTGGGTTTTCGGATTTCGCAGGGGGGAGAAATTTATTGGCCTACTTGGGTCTTTTGGAAAATAAAAATAATGAGAGAATGGGGGTTCAAGGCTTGTGATAACTGATTTGTATAGAATGATTCAATA   9558
ACTGGAACTATATGAGTGTAGGTCATAATAAAGCTAATGTGGGTAGACATTTTGATTAAAATTCGTCGAC 9628
                                                                 3'
```

TABLE 1

Sequence comparisons of H, D, L

| | DNA | | Protein |
|---|---|---|---|
| H/D | 56 | [5] | 48 |
| D/L | 57 | [5] | 49 |
| H/L | 60 | [2] | 53 |

Numbers are given in percent sequence homology.

The number in [ ] indicates the number of gaps inserted to maximize the homology.

# CHAPTER 4

# A transposable element which splits the promoter region inactivates a *Drosophila* cuticle gene

MICHAEL SNYDER, DEBORAH KIMBRELL[*], MICHAEL HUNKAPILLER, JAMES FRISTROM[*] AND NORMAN DAVIDSON[+]

Division of Biology and [+]Department of Chemistry, California Institute of Technology, Pasadena, California 91125

[*]Department of Genetics
University of California
Berkeley, California 94720

Draft for submission to P.N.A.S.

**ABSTRACT** Two mutations which affect larval cuticle gene expression in the 2/3 Drosophila melanogaster strain have been investigated. We demonstrate that this strain makes a variant cuticle protein 2, called CP2V, of altered electrophoretic mobility on two-dimensional gels. It also fails to synthesize any detectable cuticle protein 3 (CP3). The other major cuticle proteins are still present. Protein and DNA sequencing indicate that point mutations which cause two amino acid substitutions are responsible for the change in electrophoretic mobility of CP2. The mutation abolishing the expression of CP3 was found to be a 7.3 kb DNA insertion located within the TATA box region of this gene, at -31 bp from the mRNA start site. This DNA insertion, called Gulliver*, belongs to a conserved family of repeated DNA elements which have characteristics similar to that of previously characterized Drosophila transposable elements. Gulliver elements are repeated approximately 50-fold in the haploid genome and exhibit restriction fragment length polymorphisms between Canton S and 2/3 Drosophila strains. Sequence analysis indicates that Gulliver contains 266 bp direct repeats at its termini and has caused duplication of 4 bp of target DNA sequence, TATA, in the CP3 gene insertion. Thus, insertion of a transposable element into the putative promoter region of the CP3 gene is evidently responsible for inactivating CP3 gene expression.

*Previously called Zup.

The fusion of genetics and biochemistry has played a decisive role in understanding the mechanisms of gene expression in procaryotes. The use of mutants has allowed both the delineation of sequences required for a particular gene's expression and identification of the proteins that interact with these sequences. In eucaryotes the study of mutants at a molecular level has only recently begun.

We have been studying the larval cuticle genes of _Drosophila_ as a model system for understanding the regulation of gene expression in eucaryotes. Five major cuticle proteins are synthesized and secreted by the epidermal cells of late third instar larvae (1). Genes for four of the five major proteins are clustered in a small (7.9 kb) segment of the _Drosophila_ genome located at 44D on the second chromosome. The four genes in this gene cluster are related in sequence; and although there is a homologous pseudogene in this cluster, no other closely related genes exist in the _Drosophila_ genome (2,3). The 44D cuticle genes are suitable for the study of mutants because the region encoding these genes has been cloned and the gene cluster almost entirely sequenced (2,3).

We have investigated naturally occurring mutations which affect _Drosophila_ cuticle gene expression. One _Drosophila_ _melanogaster_ strain, called 2/3 (1), fails to make two of the five major cuticle proteins, but instead makes one new protein secreted into the cuticle, called CP2V. A molecular characterization of cuticle genes in this strain is reported below, including the finding of an insertion of a transposable element in the TATA box region of an unexpressed cuticle gene.

## MATERIALS AND METHODS

**Materials, Clones, and Fly Stocks.** Lambda clones and pBR322 subclones containing 44D Canton S _Drosophila_ cuticle genes have been described (2,3). The 2/3 fly strain is a naturally occurring _Drosophila_ _melanogaster_ stock isolated by Fristrom _et_ _al._ (1). Restriction endonucleases were purchased from New England Biolabs or were prepared by Maria Alonso.

**Construction and Screening 2/3 Drosophila Libraries.** A library of 2/3 DNA was cloned into the lambda vector λL47 (4) according to standard protocols (5). Six 50 μl aliquots each containing 15 μg of 2/3 DNA were partially digested at 37°C for 7 minutes with 0.11, 0.22, 0.45, 0.9, 1.5 or 2.4 units of MboI. The DNAs were pooled and size-fractionated on a 10-40% sucrose gradient (5), and 1.5 μg of 15-20 kb sized DNA was treated with calf alkaline phosphatase and ligated into 3 μg of λL47 arms, which were generously provided by N. Davis Hershey. Packaging was carried out according to Mullins et al. (6), with an efficiency of 0.5-1 X $10^6$ phage/μg insert DNA. 30,000 plaques were screened with $^{32}$P-labeled probes to genes II ($10^7$ DPM) pCPII-7, III (3 x $10^7$ DPM) pCPIII-9 and IV ($10^7$ DPM), pCPIV-8, as previously described (2). Five positive phage were chosen for further study (Fig. 2).

**Nucleic Acid Preparations.** Phage and plasmid DNAs were grown as described (2). The 5.0 kb XhoI fragment of λ1 was purified from a 0.7% agarose gel using hydroxyapatite according to (7). The 2.5 kb EcoRI fragment of λ1 was prepared similarly and subcloned into the EcoRI site of pBR322 as described (7). The resulting subclone is designated pZ-2.5.

**Genome Southerns.** 3 μg of genomic DNA was digested with restriction endonucleases and analyzed on a 0.7% agarose gel (2). Gel blots were performed using 0.45 μm nitrocellulose sheets (Millipore) and probed as described in text. Hybridizations in 50% formamide, 1 M NaCl and 10% (w/v) dextran sulfate and washes were as described previously (2). Final stringent washes were performed on 0.3X SSC (1X SSC = 0.15 M NaCl, 0.015 M Na citrate) at 65°C. In the blot in Fig. 5, 3 μg of high molecular weight calf thymus DNA plus genomic equivalents of 1, 3, 8 and 20 copies of the gel-isolated 5.0 kb XhoI fragment of λ1 was treated with XhoI plus EcoRI and analyzed in adjacent lanes as standards.

**DNA Sequencing.** Sequencing was carried out according to the strategy shown in Fig. 4A. DNA fragments were $^{32}$P labeled at their 3' ends using the large fragment of E. coli DNA polymerase I and sequenced according to the procedures of Maxam and Gilbert (8) as previously described (3). All sequences to the left of Gulliver (Fig. 4A) were determined from λ1 and to the right of Gulliver from the subclone pZ-2.5.

DNA fragment and subclones were $^{32}P$ labeled by nick translation as described by Maniatis et al. (9).

**Protein Isolation, Sequencing and Two-Dimensional Analysis.** Total cuticle proteins were prepared from late third instar larvae and 5-10 μg were analyzed on two-dimensional O'Farrell gels as previously described (2).

## RESULTS

**The 2/3 Drosophila Strain Has Two Genetic Differences Affecting Cuticle Protein Synthesis.** In most Drosophila melanogaster strains thus far examined, five major cuticle proteins and a number of minor proteins are extracted from purified cuticles of third instar larvae (1). The major proteins are called CP1 through CP5; their electrophoretic mobility pattern on two-dimensional gels is shown in Fig. 1A. However, a naturally occurring Drosophila melanogaster strain, called 2/3, has been isolated by Fristrom et al. (1). Cuticle proteins isolated from third instar larvae of this strain show the two-dimensional gel pattern presented in Fig. 1B. Only three of the major cuticle proteins CP1, CP4, and CP5 are found; both CP2 and CP3 are undetectable. Instead there is one new protein of altered electrophoretic mobility; its pI (5.5) and apparent molecular weight are less than that of CP2. This new protein is renamed CP2V, because, as shown below, it is actually a variant of CP2. [The protein was previously called 2/3 in (1) and 3V in (2).] We have sequenced 55 amino terminal residues of CP2V. Comparison with the sequence previously determined for CP2 (3) shows 54/55 residues match perfectly; one amino acid polymorphism occurs at position 18. The CP2V sequence does not match any other cuticle protein sequence. Hence for the 2/3 Drosophila strain the following questions may be asked: 1) what is the nature of the mutation affecting the electrophoretic mobility of CP2; and, more interestingly, 2) what is the nature of mutation such that no CP3 is found?

We find that in addition to lacking CP3 protein, the 2/3 Drosophila strain has reduced amounts (less than 10%), if any, of translatable CP3 RNA in late third instar larvae. A detailed analysis of the RNA levels throughout third instar larval development for each of the cuticle proteins CP1, CP2, CP3 and CP4 in Oregon R and 2/3 Drosophila melanogaster strains will be presented later (Kimbrell and Fristrom, unpublished).

**A 7.3 kb DNA Insertion is Located at the 5' End of the Unexpressed Gene III.** The genes encoding CP1 through CP4, denoted genes I through IV, respectively, have been cloned (2,3). All four genes are clustered within 7.9 kb of DNA; their organization is presented in Fig. 2. A pseudogene, ψI, also lies within this cluster (3). Each of the cuticle genes contains a short (56-64 bp) intron interrupting the signal peptide coding sequence.

In order to understand the nature of the mutations affecting CP2 and CP3 we have investigated the genomic organization of the 44D cuticle genes in the 2/3 Drosophila strain. Subcloned probes derived from the wild-type clones for each of genes I, II, III and IV (see Fig. 2) were used to probe gel blots of 2/3 genomic DNA digested with a variety of restriction endonucleases. Seven different DNA digests were probed with gene I and II probes, and 14 different DNA digests were probed with III and IV probes. In many cases Canton S DNA was digested and analyzed in adjacent lanes for comparison. Examples of the genomic blots are shown in Fig. 3, and the derived restriction map is shown in Fig. 2. The results indicate that the region within and around genes I, II and IV are identical in Canton S and 2/3 DNAs at a resolution of 100 bp. However, a 7.3 kb DNA insertion is located within 50 bp of the mRNA start site of gene III. This DNA insertion we have called Gulliver. (Compare lanes Ava, Bgl+H3, RI+H3 of 2/3 and CS DNAs, Fig. 3III.)

**Isolation of 2/3 DNA Clones Containing 44D Cuticle Genes.** We further investigated the insertion mutation near or within gene III and the mutation affecting gene II by cloning the 44D cuticle gene region of the 2/3 Drosophila melanogaster genome. Libraries of 2/3 DNA were constructed in the lambda vector λL47 (4) and in the plasmid vector pBR322. A series of overlapping clones was isolated with inserts containing different parts of the 2/3 cuticle gene region. One clone, called λ2/3LCP1 (λ1), which has been studied in detail, contains an entire copy of Gulliver and its flanking sequences.

**The Gulliver Insertion is in the TATA Box Region of the Unexpressed Gene III.** In order to determine the position of the Gulliver insertion relative to gene III we sequenced the junctions of the Gulliver insertion and gene III DNA. The sequencing strategy and result are presented in Fig. 4. Novel DNA sequences begin (Fig. 4b) at position -32 bp from the mRNA start site, immediately adjacent to the $\overset{-31}{\text{T}}\text{A}\overset{-29}{\text{T}}\text{ATAAA}$ sequence (TATA box) of gene III. [The TATA box is placed to start at -31 rather than -29 based on its homology to the other cuticle genes and also by its position relative to the mRNA start site (3).] At the left junction of the Gulliver insertion we find another copy of part of the TATA box, TATA, plus the expected upstream flanking sequence. We conclude that the Gulliver element has inserted within or immediately adjacent to the TATA box and has duplicated 4 bp of TATA box sequence (see below). In the remainder of gene III we find no other nucleotide changes in either the mRNA coding sequences (489 bp) or in the upstream flanking sequence (119 bp).

In order to understand the nature of the mutation affecting the electrophoretic mobility of CP2 in the 2/3 strain we have sequenced gene II (Fig. 4). From the DNA sequence we predict only two amino acid substitutions in the protein coding region. One of these is a substitution of leucine for serine and the other is a replacement of serine for arginine. These substitutions are consistent with the observed pI difference between CP2V and CP2. Only one other nucleotide change is found in the entire gene II region; it lies in the 3' untranslated region (Fig. 4C).

From our DNA sequencing results the overall nucleotide polymorphism between 2/3 and Canton S DNA was determined to be 0.16% (3/1844 bp) excluding the Gulliver insertion. [3/∼985 are in mRNA coding regions; 0/∼859 are in non-mRNA coding regions.] The remaining DNA sequence of gene II and III regions not shown in Fig. 4 can be found in (3).

**Gulliver Has Characteristics Similar to That of Other Transposable Elements.** The exact termini of Gulliver are defined by the DNA sequences found at the insertion junctions. DNA sequencing shows that 266 bp direct repeats are found at the two ends of Gulliver (Fig. 4). The termini of the 266 bp repeats contain short, imperfect, inverted repeats of 7 bp, which begin with the sequence AGT at the left end (Fig. 4). As discussed above, a 4 bp sequence, TATA, flanking Gulliver DNA is duplicated. These features 1) long direct repeats (276 to 571 bp) (which begin with the sequence $^{T}_{A}$GT or AC$^{A}_{T}$ in other instances studied so far), 2) short invert repeats (14 bp or less) at the direct repeat termini and 3) the 4-5 bp duplication of target DNA sequence, are all properties common to many *Drosophila* transposable elements [see review (12)].

To characterize further the nature of Gulliver we determined its repetition frequency in the *Drosophila melanogaster* genome. An internal 5.0 kb *XhoI* fragment of Gulliver was $^{32}$P-labeled and used to probe gel blots of 2/3 genomic DNA digested with *EcoRI* and with *EcoRI* plus *XhoI*. As shown in Fig. 5 the element is repeated and most of its members contain the predicted internal *XhoI* fragment. Comparison with standards (Fig. 5, using different exposure times) reveals that there are approximately 50 copies of Gulliver in the *Drosophila melanogaster* haploid genome.

Comparison of 2/3 and Canton S *EcoRI* digested DNAs which were probed with the internal *XhoI* fragment show restriction fragment length polymorphism of Gulliver-homologous sequences (Fig. 5). The repetitive nature and restriction fragment length polymorphism exhibited by Gulliver sequences are also characteristic of Drosophila transposable DNA elements (12).

## DISCUSSION

The data presented above show that two types of mutations affect cuticle protein synthesis in the Drosophila melanogaster 2/3 strain. Two amino acid substitutions in gene II are evidently responsible for the production of CP2V protein of altered electrophoretic mobility as compared to CP2. The substitution of a serine for an arginine is consistent with the observed pI shift of CP2V from CP2. However, the decrease of 10% in SDS gel electrophoretic mobility of CP2V relative to CP2 is unexpected since our sequence data predict that the two proteins have the same number of amino acid residues. Since no post-translational modification (other than signal peptide processing) is known for these proteins including glycosylation, acetylation and phosphorylation (3), we suggest that the amino acid substitutions found in CP2V are responsible for the apparent mobility differences observed. Previous data have shown that proteins of similar size which differ only in charge sometimes display an apparent molecular weight difference on SDS gels (11). However, we cannot preclude the occurrence of unknown post-translational modifications. Such modifications must lie in the carboxy terminal half of the protein, as the amino terminal half of the protein has been sequenced and no modified residues found.

The mutation affecting the expression of the CP3 gene appears to be the Gulliver DNA insertion immediately adjacent to the TATATAAA sequence of gene III, and has duplicated 4 bp, TATA of this sequence. This mutation is probably responsible for inactivating the expression of gene III, because no other mutations within the gene III mRNA coding or upstream flanking sequences exist. In addition, data from other eucaryotic genes have indicated that a TATA box sequence plus sequences upstream of the TATA box are necessary for efficient transcription in vivo (for example, 13,14).

Gulliver is a member of copia-like middle repetitive Drosophila DNA, many families of which, including copia, 297, 412, mdg 1, mdg 3, roo, gypsy and others, have been well studied (reviewed in 12; roo from 15). Gulliver has a structure and properties similar to these transposable elements; 1) direct repeats at both ends; 2) short inverted repeats at the direct repeat termini; 3) restriction fragment length polymorphism of family members in two Drosophila strains; and 4) a small duplication of target DNA sequence. Examples of spontaneous mutations caused by transposable element insertions have been frequently noted in the Drosophila white (16-18) and bithorax loci (19) as well as in yeast loci (20-23). Their occurrence in vertebrates has yet to be demonstrated.

Recently, another example of a transposable element, 297, which inserted into the sequence TATATA in front of a Drosophila histone gene has been reported (24, 25). Since Gulliver belongs to a different family of copia-like elements as judged by restriction map and sequence analysis, two such events suggest that insertion of such elements into the TATATA sequence may be more general, and have important biological implications in abolishing or altering the manner in which genes are developmentally regulated. We note that the mechanism of insertion of 297 and Gulliver must differ in at least one respect: 297 duplicates 5 bp of DNA sequence (12, 24), and Gulliver duplicates 4 bp (Fig. 4).

Note that in the 2/3 Drosophila strain the new CP2V protein resembles the missing CP3 in that it is less basic than CP2. Furthermore, no electrophoretic variants affecting only CP2 or CP3 have yet been found. However, two variant strains having the 2/3 phenotype described here have been isolated independently, from widely separated regions of the world. Perhaps both CP2 and CP3 are required for larval cuticle formation and Drosophila melanogaster strains with mutations that affect one gene are selected against unless they contain a compensating mutation in the other gene.

## Acknowledgements

We thank Dr. N. Davis Hershey for generously providing λL47 arms, Dr. N. Davis Hershey and Mark Garfinkel for their criticisms of the manuscript and Dr. Lee Hood in whose laboratory the protein sequencing was performed. This research was supported by grants GM 20927 (ND), GM 0695 (MH), and ___________ (JF). MPS was supported by the training grant GM 07616.

## REFERENCES

1. Fristrom, J. W., Hill, R. J. & Watt, F. (1978) Biochemistry **19**, 3917-3924.
2. Snyder, M., Hirsh, J. & Davidson, N. (1981) Cell **25**, 165-177.
3. Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. & Davidson, N. (1982) Cell **29**, 1027-1040.
4. Loenen, W. A. M. & Brammar, W. J. (1980) Gene **10**, 249-259.
5. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) Cell **15**, 687-701.
6. Mullins, J. I., Casey, J. W., Nicolson, M. O., Burck, K. B. & Davidson, N. (1981) J. Virology **38**, 688-703.
7. Fyrberg, E. A., Kindle, K. L., Davidson, N. & Sodja, A. (1980) Cell **19**, 365-378.
8. Maxam, A. M. & Gilbert, W. (1980) in Methods in Enzymology, 65, pp. 499-559.
9. Maniatis, T., Jeffrey, A. & Kleid, D. G. (1975) Proc. Natl. Acad. Sci. USA **72**, 1184-1188.
10. Hunkapiller, M. & Hood, L. E. (1980) Science **207**, 523-525.
11. Papkoff, J., Verma, I. M. & Hunter, T. (1982) Cell **29**, 417-426.
12. Spradling, A. C. & Rubin, G. M. (1981) Ann. Rev. Genet. **15**, 219-264.
13. McKnight, S. L., Gavis, E. R., Kingsbury, R. & Axel, R. (1981) Cell **25**, 385-398.
14. Grosschedl, R. & Birnstiel, M. L. (1980) Proc. Natl. Acad. Sci. USA **77**, 7102-7106.
15. Meyerowitz, E. M. & Hogness, D. S. (1982) Cell **28**, 165-176.
16. Bingham, P. M., Levis, R. & Rubin, G. M. (1981) Cell **25**, 693-704.
17. Rubin, G. M., Kidwell, M. G. & Bingham, P. M. (1982) Cell **29**, 987-994.
18. Zachar, Z. & Bingham, P. (1982) Cell, in press.
19. Hogness, D. S., Saint, R. B., Akam, M. E., Goldschmidt-Clermont, M. & Beachy, P. (1982) J. Cell. Biochem **6**, 263.
20. Chaleff, D. T. & Fink, G. R. (1980) Cell **21**, 227-237.
21. Roeder, G. S. & Fink, G. R. (1980) Cell **21**, 239-249.

22. Errede, B., Cardillo, T. S., Sherman, F., Dubois, E., Deschamps, J. & Wiame, J.-M. (1980) Cell **25**, 427-436.

23. Williamson, V. M., Young, E. T. & Ciriacy, M. (1981) Cell **23**, 605-614.

24. Ikenaga, H. & Saigo, K. (1982) Proc. Natl. Acad. Sci. USA **79**, 4143-4147.

25. Goldberg, M. (1979) Ph.D. Thesis, Stanford University.

Fig. 1. Comparison of Third Instar Cuticle Proteins of Canton S and 2/3 *Drosophila* Strains. Coomassie stained two-dimensional O'Farrell gels (2). Most wild-type *Drosophila melanogaster* strains contain cuticle proteins CP1 through CP5 (1). The 2/3 *Drosophila melanogaster* strain lacks detectable amounts of CP2 and CP3 in purified cuticles, but makes a different protein called CP2V. Molecular weights are given in kd. The molecular weights of CP1 and CP2 have been revised slightly based on predictions from DNA sequencing results (3).

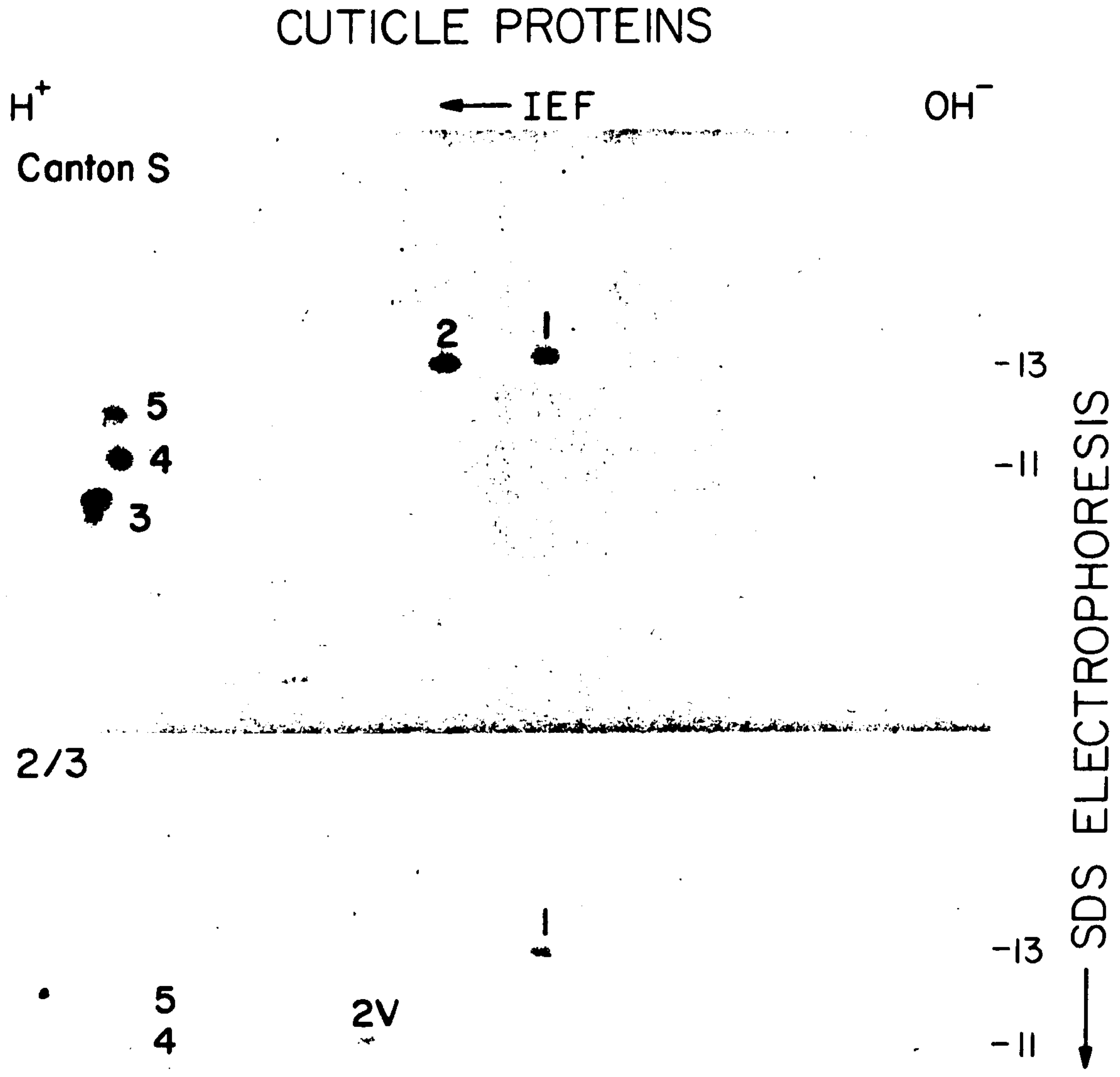

CUTICLE PROTEINS
H+
IEF
OH−
Canton S
2
1
-13
5
4
-11
3
SDS ELECTROPHORESIS
2/3
1
-13
5
4
2V
-11

Fig. 2. The Organization of 44D Cuticle Genes. The upper diagram shows a restriction map of the cloned region of the Canton S region. Genes I through IV encode CP1 through CP4, respectively (2,3). ψ I is a putative Drosophila cuticle pseudogene. Arrows beneath the genes indicate the 5' to 3' direction of transcription (3). The subclones pCPI-11, pCPII-7, pCPIII-9, pCPIV-8 are subclones of Canton S DNA encoding genes I through IV, respectively, and contain the regions indicated. In 2/3 DNA the restriction map is identical except for the presence of the 7.3 kb DNA insertion, Gulliver.

Beneath the map are shown overlapping cloned inserts of 2/3 DNA. λ1 through λ5 are abbreviations for λDm2/3 LCP1 through 5, respectively. All sites from the BamHI site 1.3 kb to the right of gene IV to the HindIII site 1.0 kb to the left of gene I have been mapped by genomic Southerns blotting experiments using 2/3 DNA (see text and Fig. 3). Sites outside this region were mapped only on cloned DNA inserts. The AvaI map is incomplete; only sites in the I through IV region are shown and closely spaced AvaI sites have not been indicated. Note also that AvaI also cleaves at XhoI sites (X). Key: R = EcoRI; B = BamHI; O = BglII; S = SalI; C = SacI; H = HindIII; A = AvaI; K = KpnI; X = XhoI.

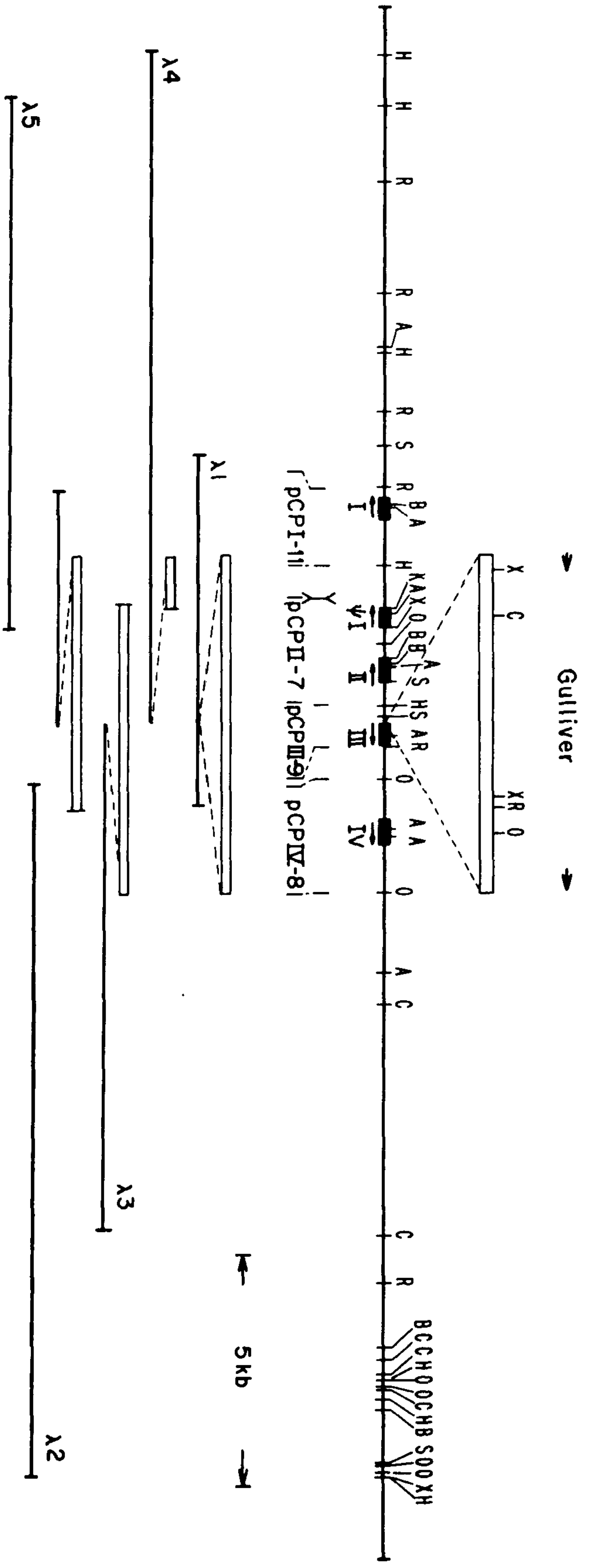

Gulliver
X
C
XR
O
H
H
R
R
A
H
R
S
R
B A
H
KAXO BE
A
S
HS AR
O
A A
O
A
C
C
R
BCCHOOCHB
SOOXH
I
ψI
II
III
IV
pCPI-11
pCPII-7
pCPIII-9
pCPIV-8
λ1
λ2
λ3
λ4
λ5
5 kb

Fig. 3. Genome Southerns of 2/3 DNA Probes with I: pCPI-11, II) pCPII-7, III) pCPIII-9 and IV) pCPIV-8. Enzymes used are as indicated. Note the gene I and II fragments crosshybridize in the mRNA coding regions as do the III and IV fragments. We have indicated this cross-hybridization by an X when the homologous genes are separated on different sized restriction fragments. For comparison Canton S (CS) and 2/3 DNAs are shown. Additional gel blots of 2/3 DNA digested with other combinations of restriction endonucleases and probed with pCPIII-9 and pCPIV-8 have been performed but are not shown. The ordinate indicates the size in kb of AvaI hybridizing bands.

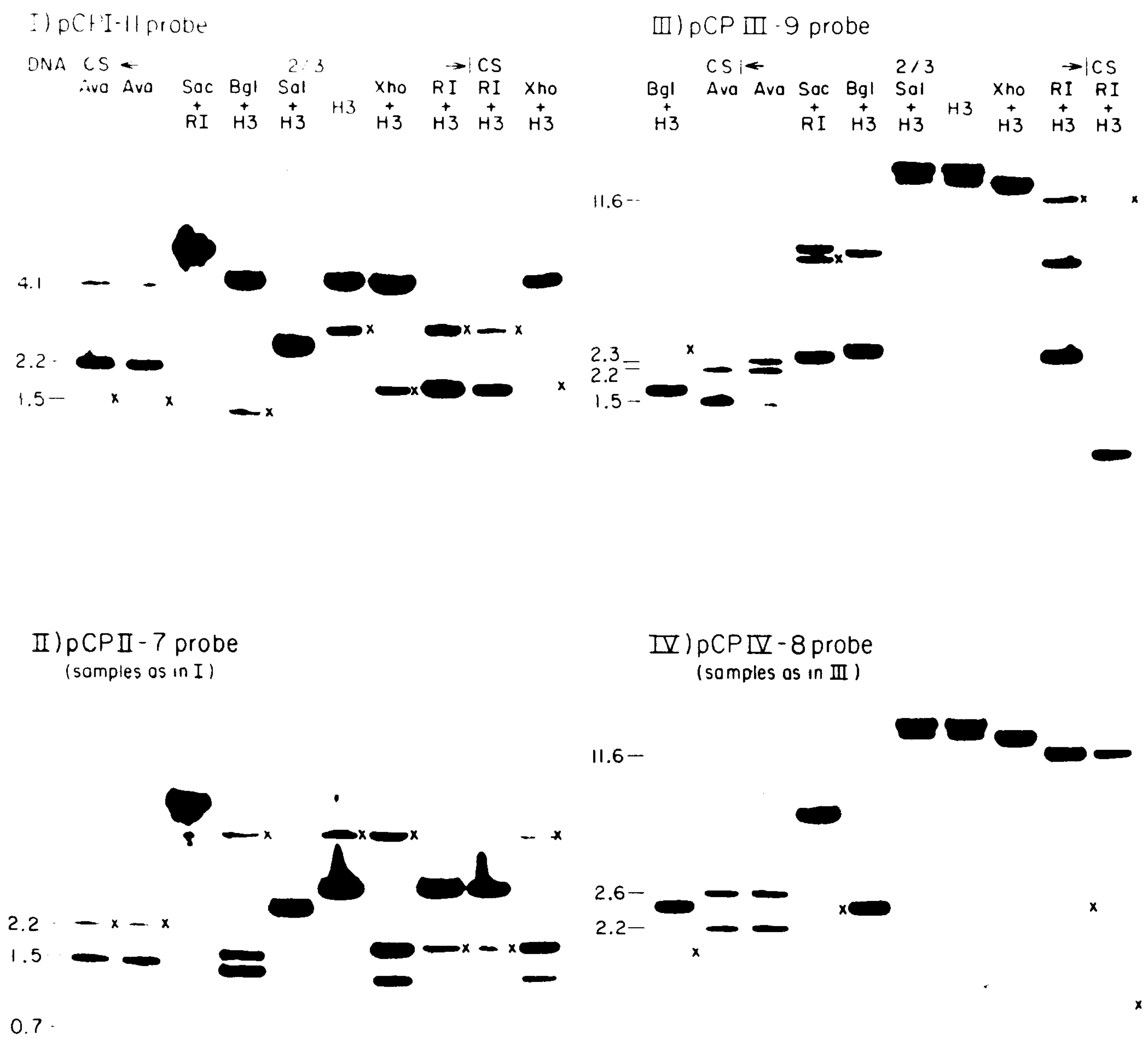
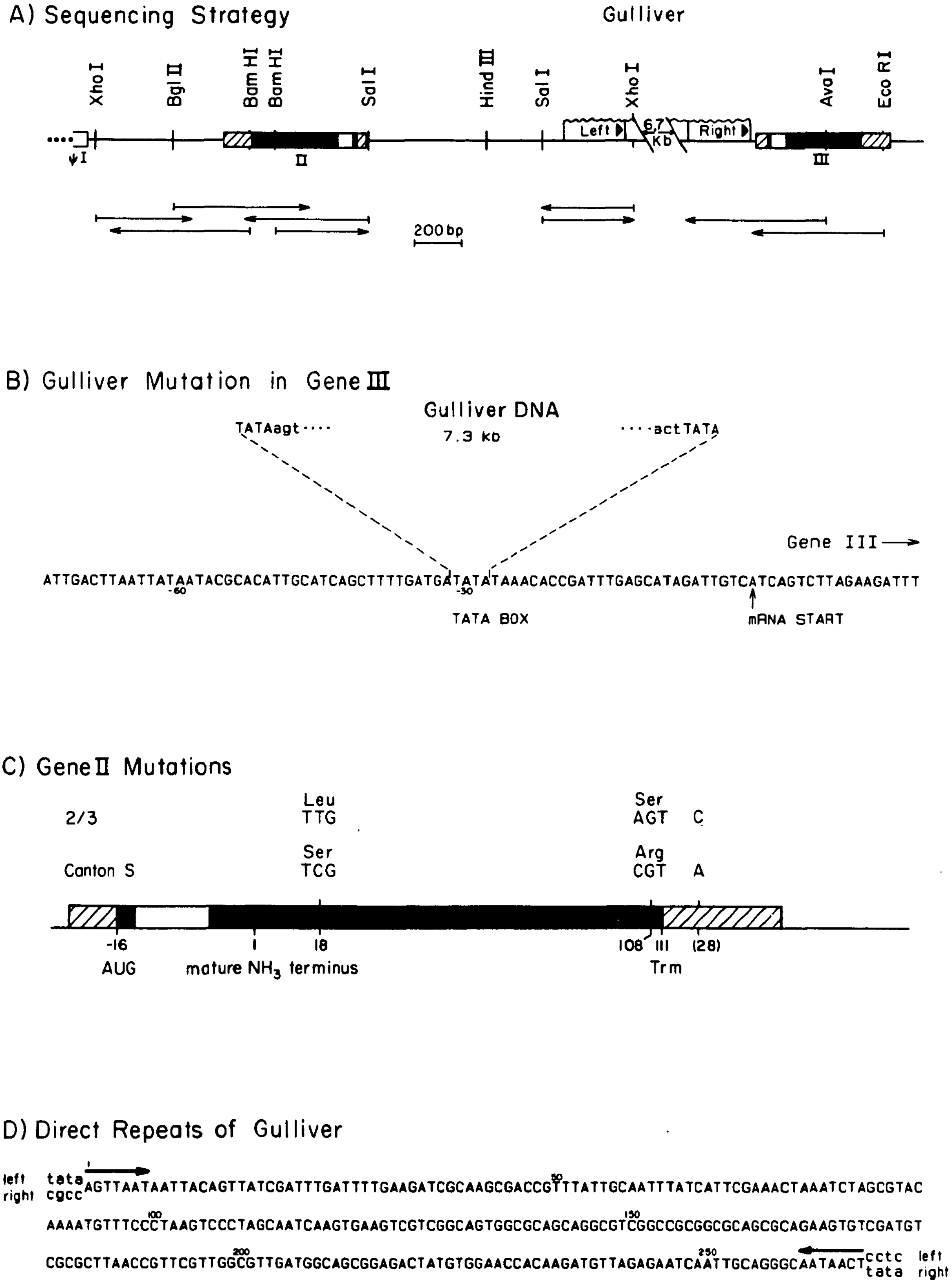

I) pCPI-II probe
DNA CS ←
Ava Ava Sac + RI Bgl + H3 Sal + H3 H3 Xho + H3 RI + H3 RI + H3 Xho + H3
2/3
→| CS
4.1
2.2
1.5
x
III) pCP III-9 probe
CS |←
Bgl + H3 Ava Ava Sac + RI Bgl + H3 Sal + H3 H3 Xho + H3 RI + H3 RI + H3
2/3
→| CS
11.6
2.3
2.2
1.5
II) pCP II-7 probe
(samples as in I)
2.2
1.5
0.7
IV) pCP IV-8 probe
(samples as in III)
11.6
2.6
2.2

Fig. 4. Sequencing Gulliver Insertion Junctions and Genes II and III. (A) Sequencing strategy. The diagram shows gene II and III and the Gulliver insertion regions. The arrows indicate the direction and extent of DNA sequencing of fragments $^{32}$P-labeled at their 3' termini. DNA sequencing was according to Maxam and Gilbert (8). Only restriction sites pertinent to the sequencing are indicated. Key: solid box-protein coding region; diagonal box-mRNA untranslated region; open box-intron. (B) The sequence of the gene III and Gulliver insertion regions. The wild-type sequence is shown. The insertion in the 2/3 DNA is indicated above the wild-type sequence. Note the sequence, TATA, of the TATA box is on both sides of the insertion. No other nucleotide changes in the gene III region have been found; the complete sequence from the SalI site (-119) to the EcoRI site at the 3' end of gene III is identical to the Canton S sequence (3). (C) Sequence of the gene II region. The only three differences between wild-type and 2/3 DNA are indicated. The remainder of the DNA sequence which extends to the XhoI site, (Fig. 4A) is identical in 2/3 and Canton S DNAs (see 3 for complete sequence). The first 55 amino acid residues of CP2V were also determined and are identical to that predicted from the DNA sequence. Units on the axis indicate codon number; codon No. 1 encodes the amino terminal residue of the mature protein. (28) indicates a nucleotide substitution 28 bp away from the termination codon.
(D) The terminal repeats of Gulliver. Large letters indicate Gulliver direct repeats. Small letters indicate the flanking sequences. Left and right indicate the left and right repeats as shown in panel a.



A) Sequencing Strategy

B) Gulliver Mutation in Gene III

ATTGACTTAATTATAATACGCACATTGCATCAGCTTTTGATGATATATAAACACCGATTTGAGCATAGATTGTCATCAGTCTTAGAAGATTT

C) Gene II Mutations

D) Direct Repeats of Gulliver

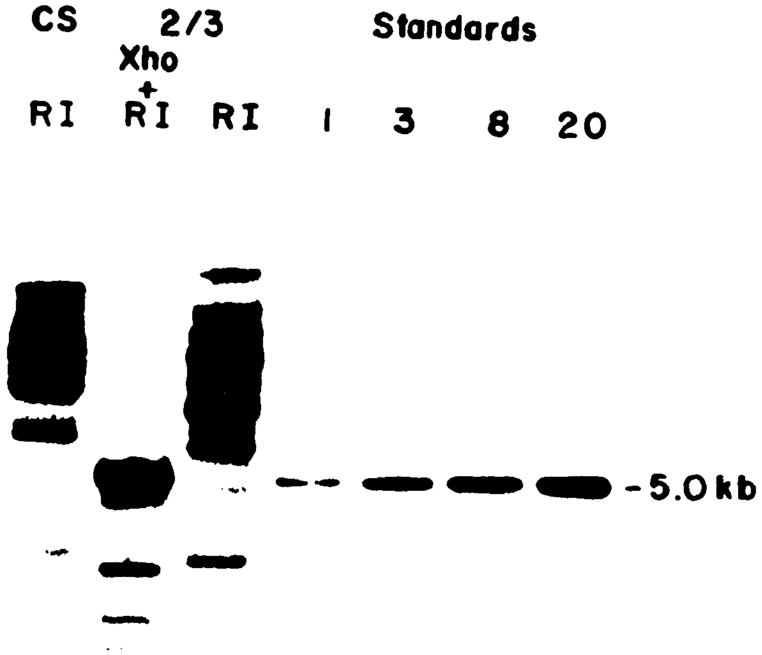
left tata
right cgcc AGTTAATAATTACAGTTATCGATTTGATTTTGAAGATCGCAAGCGACCGTTTATTGCAATTTATCATTCGAAACTAAATCTAGCGTAC

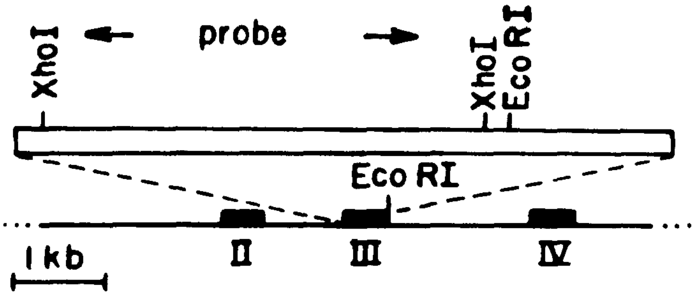
AAAATGTTTCCCTAAGTCCCTAGCAATCAAGTGAAGTCGTCGGCAGTGGCGCAGCAGGCGTCGGCCGCGGCGCAGCGCAGAAGTGTCGATGT

CGCGCTTAACCGTTCGTTGGCGTTGATGGCAGCGGAGACTATGTGGAACCACAAGATGTTAGAGAATCAATTGCAGGGCAATAACT cctc left
tata right

Fig. 5. Representation of Gulliver Sequences in 2/3 and Canton S DNAs. 3 μg of 2/3 and Canton S (CS) DNAs were digested with restriction endonucleases and analyzed on a 0.7% agarose gel. As standards 3 μg of calf thymus DNA mixed with genomic equivalents of 1, 3, 8 and 20 copies of an internal 5.0 kb XhoI fragment isolated from λ1 were digested with XhoI + EcoRI and analyzed on adjacent lanes. Gel blots were prepared and probed with the internal 5.0 kb XhoI fragment of Gulliver.

Genome Southerns





# APPENDIX

# APPENDIX 1

## Cuticle Protein Sequences

Figure A1. Summary of eight third instar larval cuticle protein sequences. CP1 through 4 sequences have been previously described (Snyder et al., 1982). Sequences of the fifth major cuticle protein, CP5, and three minor proteins CP6, CP2a, CPX are derived by amino terminal protein sequencing and are incomplete (Snyder, Silvert, Hunkapiller, Fristrom and Davidson, unpublished). CPX is a minor third instar larval protein of unknown identity.

Figure A2. Relative Hydrophobicity Content as Deduced from the Protein Sequence of CP2. The ordinate depicts a relative hydrophobicity value determined from analyzing nine amino acid residues centered on the amino acid residue indicated on the abscissa. The hydrophobicity value was calculated by Dr. Russ Doolittle according to Kyte and Doolittle (1982). Regions that fall substantially above the dotted line (26) are hydrophobic. All four proteins CP1 through CP4 show a similar pattern to that shown in Figure A2: the signal peptide and the region near the COOH terminus are hydrophobic.

Figure A3. Predicted secondary structure of CP1 through CP4 calculated according to the rules of Chou and Fasman (1978) by H. Lipke.

**References:**

Chou, P. Y. and Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymology 47, 45-148.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105-132.

Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. and Davidson, N. (1982). Cuticle protein genes of Drosophila: Structure, organization, and evolution of four clustered genes. Cell, 1027-1040.

Figure A1

**Cuticle Protein Sequences**

```
              Amino Terminus
                                                                                               1           5
CP1           F  K  F  V  M  I  C  A  V  L  G  L  A  V  A  N  P  P  V  P  H  S  L  G  R  S  E  D  V  H  A  D  V  L  S
CP2                       F  K  F  V  M  I  L  A  V  V  G  V  A  T  A  L  A  P  V  S  R  S  D  D  V  H  A  D  V  L  S
CP3                                               F  K  I  L  L  V  C  S  L  A  A  L  V  A  A  N  A  N  V  E  V  K  E
CP4                                               F  K  I  L  L  V  C  A  L  V  A  L  V  A  A  N  E  E  A  D  V  V  K
CPX                                                                                            N  E  E  A  D  V  V  K
CP2a                                                                                           N  E  D  A  N  V  L  R
CP5 and CP6                                                                                       N  L  A  E  I  V  R

              10             15             20             25             30             35             40
CP1           R  S  D  D  V  R  A  D  G  F  D  S  S  L  H  T  S  N  G  I  E  Q  A  A  S  G  D  A  H  G  N  I  H  [
CP2           R  S  D  D  V  R  A  D  G  F  D  S  S  L  H  T  S  N  G  I  E  Q  A  A  S  G  D  A  H  G  N  I  H  [
CP3           L  V  N  D  V  Q  P  D  G  F  V  S  K  L  V  L  D  D  G  S  A  S  S  A  T  G  D  I  H  G  N  I  D  [
CP4           L  V  N  D  V  Q  A  D  G  F  V  S  K  L  V  L  D  N  G  S  A  A  S  A  T  G  D  V  H  G  N  I  D  [
CPX           S  D  S  E  V  N  L  L  D  F  N  Y  A  Y  E  L  S  N  H  I  R  A  V  Q  T  G  A  L  K  E  H  D  N  W  V
CP2a          A  E  Q  Q  V  N  V  D  G  F  A  Y  A  V  E  L  D  N  S  V  N  V  Q  Q  K  G  D  L  N  G  E  E  [ ] W  V
CP5 and CP6   Q  V  S  D  V  E  P  D  K  I  S  S  D  V  ?  I  T  (D)(G)

                             45             50             55             60             65             70
CP1              ]  G  N  F  G  W  I  S  P  E  G  E  H  V  E  V  K  Y  V  A  N  E  N  G  Y  Q  P  S  G  A  W  I  P  T
CP2              ]  G  N  F  G  W  I  S  P  E  G  E  H  V  E  V  K  Y  V  A  N  E  N  G  Y  Q  P  S  G  A  W  I  P  T
CP3              ]  G  V  F  E  W  I  S  P  E  G  V  H  V  R  V  S  Y  K  A  D  E  N  G  Y  Q  P  Q  S  D  L  L  P  T
CP4              ]  G  V  F  E  W  V  S  P  E  G  E  H  V  R  V  S  Y  K  A  D  E  N  G  Y  Q  P  Q  S  D  L  L  P  T
CPX           V  S  G  E  Y  E  Y  V  A  P  N  G  K  T  V  K  V  V  Y  T  A  D  E  T  G  Y  N  P  K  ?  V  E  A
CP2a          V  K  G  I  Q  I  W
```

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | COOH Terminus |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 75 |  |  |  |  | 80 |  |  |  |  | 85 |  |  |  |  | 90 |  |  |  |  | 95 |  |  |  |  | 100 |  |  |
| CP1 | P | P | P | I | P | E | A | I | A | R | A | V | A | W | L | E | S | H | P | P | A | P | E | H | P | R | H | H |
| CP2 | P | P | P | I | P | E | A | I | A | R | A | V | A | W | L | E | S | H | P | P | A | P | E | H | P | R | H | H |
| CP3 | P | P | P | I | P | A | A | I | L | K | A | I | A | Y | I | E | A | N | P | S | K | N |  |  |  |  |  |  |
| CP4 | P | P | P | I | P | E | A | I | L | K | A | I | A | Y | I | Q | A | H | P | S | K | E |  |  |  |  |  |  |

- An underline indicates the amino terminus of the mature protein.
- The signal peptide is indicated in italics.
- [] no residue, inserted to maximize homology.
- ( ) indicates uncertainty of amino acid residue.

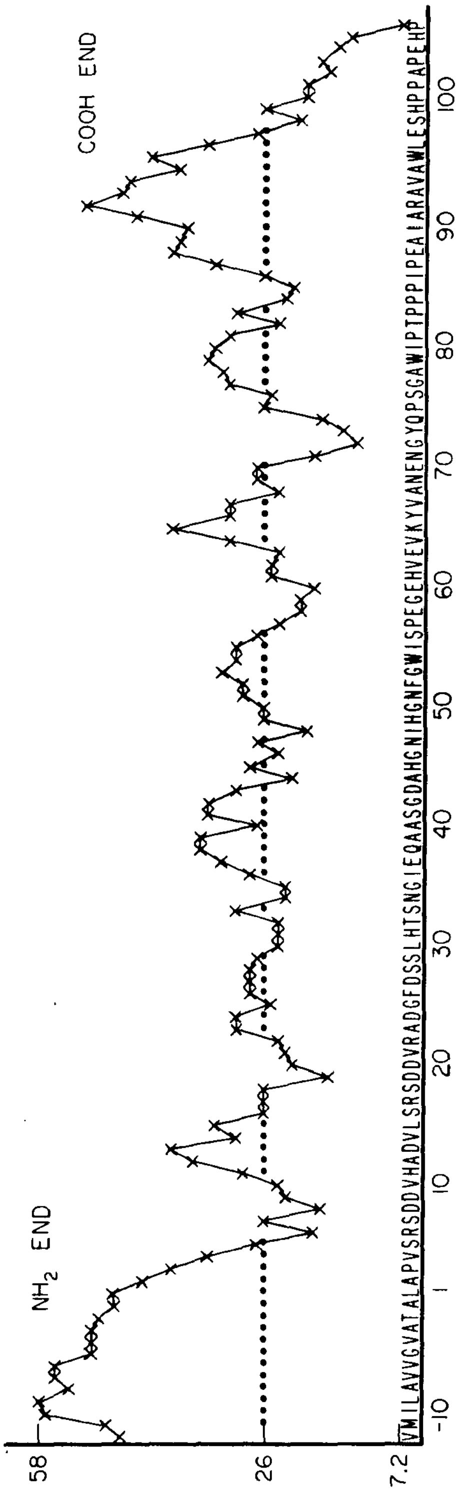

58
26
7.2
NH$_2$ END
COOH END
VMILAVVGVATALAPVSRSDDVHADVLSRSDDVRADGFDSSLHTSNGIEQAASGDAHGNIHGNFGWISPEGEHVEVKYVANENGYQPSGAWIPTPPPIPEA!ARARAVAWLESHPPAPEHP
-10
1
10
20
30
40
50
60
70
80
90
100

Figure A2

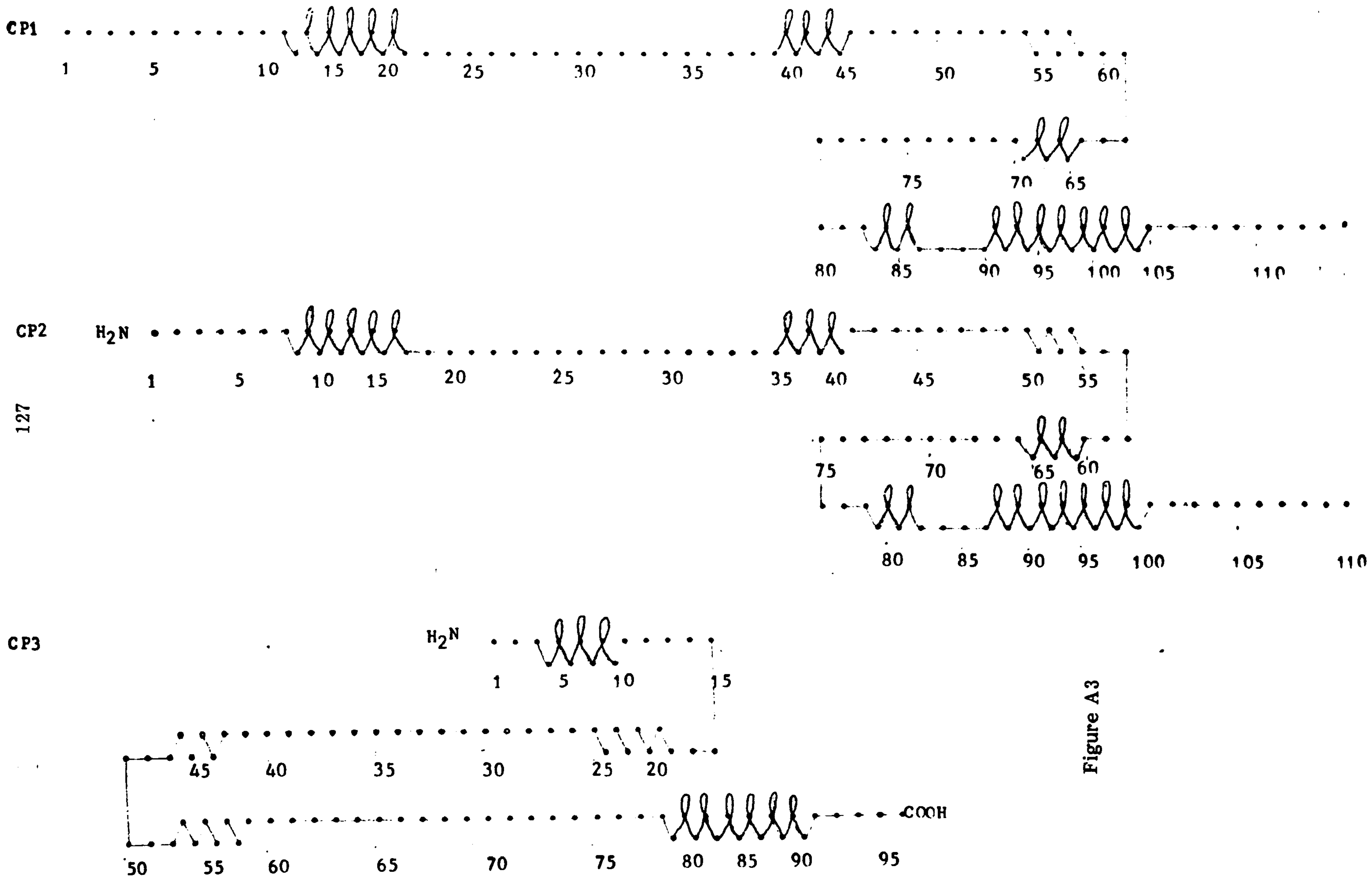

CP1
1
5
10
15
20
25
30
35
40
45
50
55
60
75
70
65
80
85
90
95
100
105
110
CP2
$H_2N$
1
5
10
15
20
25
30
35
40
45
50
55
75
70
65
60
80
85
90
95
100
105
110
CP3
$H_2N$
1
5
10
15
45
40
35
30
25
20
50
55
60
65
70
75
80
85
90
95
COOH

Figure A3

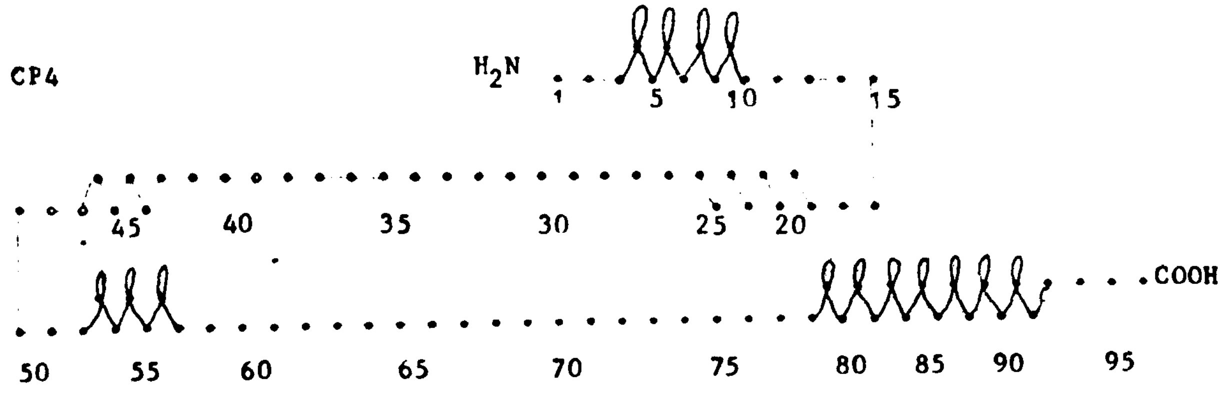


Fig. A3. Schematic diagram of the secondary structure as predicted for larval cuticle proteins in Drosophila Snyder (unpublished) using the rules of Chou and Fasman (1978). Residues are represented being in a helix ( ℓ ) or in a β sheet ( ∧ ) or coil ( · ···) conformation with bends as indicated.

# APPENDIX 2

## Second Instar Larval Cuticle Proteins

Cuticles were purified from 50-54 hr second instar Drosophila larvae, and the proteins were extracted and analyzed by native and SDS gel electrophoresis (see Snyder et al., 1981, for procedures).

Figure A2: Native (nondenaturing) gel of second instar larval cuticle proteins ($L_2$CPs). The nomenclature is that used by Chihara et al. (1982) and has been extended for previously undescribed proteins. For comparison, third instar larval cuticle proteins are analyzed in adjacent lanes. A) 1.5 µg of protein. B) 7.5 µg of protein. C) 15 µg of protein.

Figure A3: SDS gel of individual second instar larval cuticle proteins. 70 µg of second instar larval cuticle proteins were fractionated on a 40 cm native gel, individual protein bands excised, and 1-2 µg of protein reanalyzed on an SDS gel (Laemmli, 1970). The $L_2$CP2 sample is contaminated with $L_2$CP1. Molecular weight markers in kilodaltons are from bottom to top: 3, insulin, 6, bovine pancreatic trypsin inhibitor; 12-14, cytochrome C plus lysozyme; 18.4, β lactoglobulin; 25.7, α chymotrypsinogen; 43, ovalbumin.

The calculated molecular weights are presented in Table A1. Previous determinations of third instar larval cuticle protein sizes using similar gel systems have been too large for several of the proteins (Snyder et al., 1981) when compared with DNA and protein sequencing results (Snyder et al., 1982). We have therefore included estimates using the standards shown in Fig. A3 as well as estimates using third instar larval cuticle proteins as molecular weight standards (Table 1A).

### References

Chihara, C., Silvert, D. and Fristrom, J. W. (1982). The cuticle proteins of Drosophila melanogaster: Stage specificity. Dev. Biol. **89**, 379-388.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature **227**, 680-685.

Snyder, M., Hirsh, J., and Davidson, N. (1981). The cuticle genes of Drosophila: A developmentally regulated gene cluster. Cell **25**, 165-177.

Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. and Davidson, N. (1982). Cuticle protein genes of Drosophila: Structure, organization, and evolution of four clustered genes. Cell, 1027-1040.

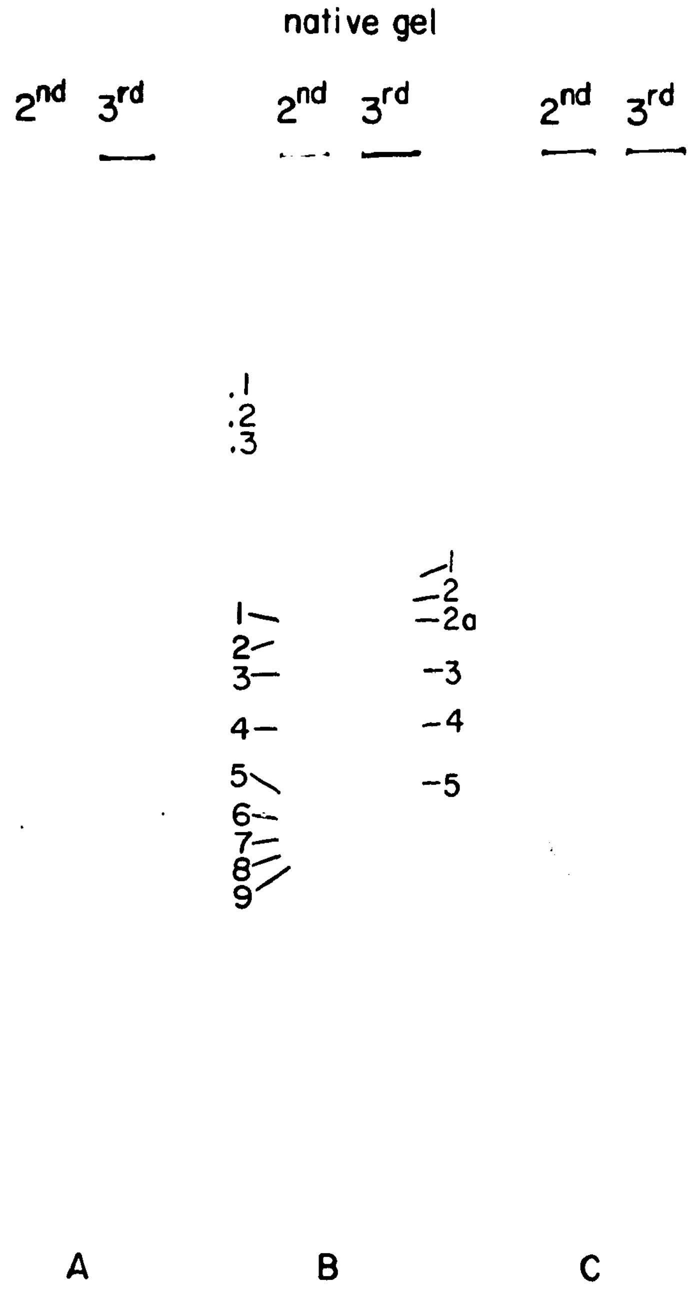


Figure A4

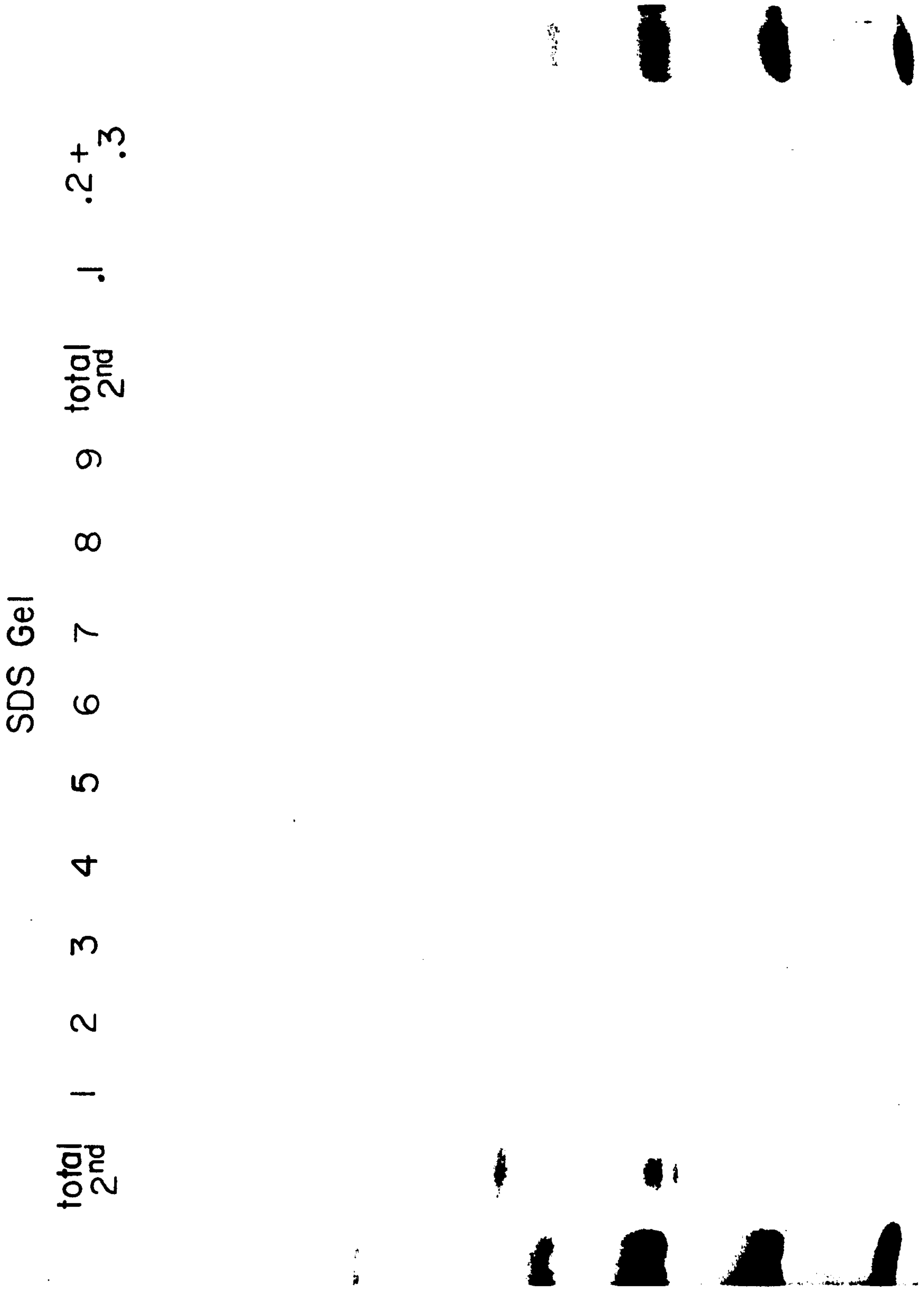


Figure A5

## TABLE A1

| Second Instar Cuticle protein | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | .1 | .2+.3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MW using Fig. A3 Standards | 12 | 15 | 11 | 10 | A) 11<br>B) 19 | A) 16<br>B) 18 | 19 | 18 | 18 | 17 | 13<br>11 |
| MW using 3rd Instar CP Standards | 10 | 12 | 10 | 9 | A) 10 | A) 12<br>B) 14 | 13 | 13 | 13 | 13 | 11 |

Units are in kilodaltons.

Two proteins are resolved in 5, 6, and 7.

# APPENDIX 3

## The Null 1 Strain of Drosophila melanogaster

The fly strain cn/br kept at the California Institute of Technology makes no detectable third instar larval cuticle protein 1 (CP1) (Chihara, Kimbrell and Fristrom, unpublished). Figure A6 shows native and SDS gels of third instar cuticle proteins of wild type (Canton S), and null 1 fly stocks. CP1 and CP2 are resolved on both of these gels (this is generally not the case for SDS gels). The results show there is no CP1, but twice as much protein at the CP2 position.

An analysis of the gene I DNA in the null 1 strain was carried out. Genomic DNA of null 1 flies was digested with many different restriction enzymes (BglII, BglII plus H3, BglII plus EcoRI, KpnI plus EcoRI, HindIII plus XhoI, KpnI plus HindIII, HindIII, HindIII plus EcoRI) and fractionated on a .7% agarose gel. Canton S DNA was digested with the same enzymes and analyzed in adjacent lanes. Gel blots were prepared and probed with pCPI-11 (Snyder et al., 1981; see Chapter 1) which contains gene I. Examples of these blots are found in Fig. A7. The result found was that the regions around gene I are identical to within 100 bp in null 1 and Canton S DNAs (Fig. A7, bottom). Since the principal difference between CP1 and CP2 is that CP1 contain one additional basic charge (and four more amino acids) than CP2, a substitution of a neutral amino acid residue for a basic residue in CP1 could account for all the observations discussed above. (The slight mobility difference of CP1 from CP2 on SDS gels (Fig. A5) may be due to the single charge difference; see Chapter 2 for precedents to this observation.)

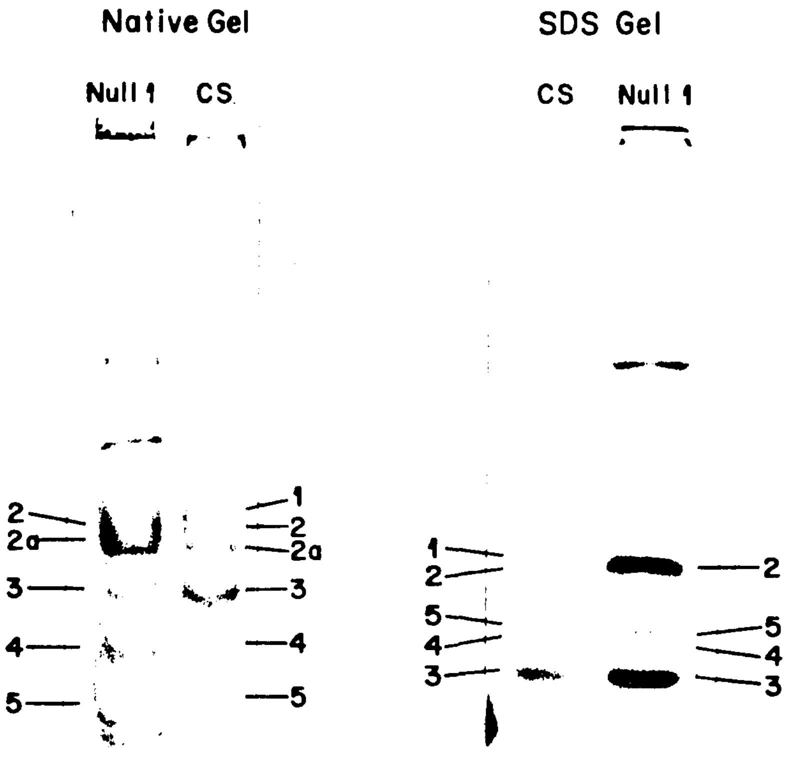

Native Gel
Null 1
CS
2
2a
3
4
5
1
2
2a
3
4
5
SDS Gel
CS
Null 1
1
2
5
4
3
2
5
4
3

Figure A6

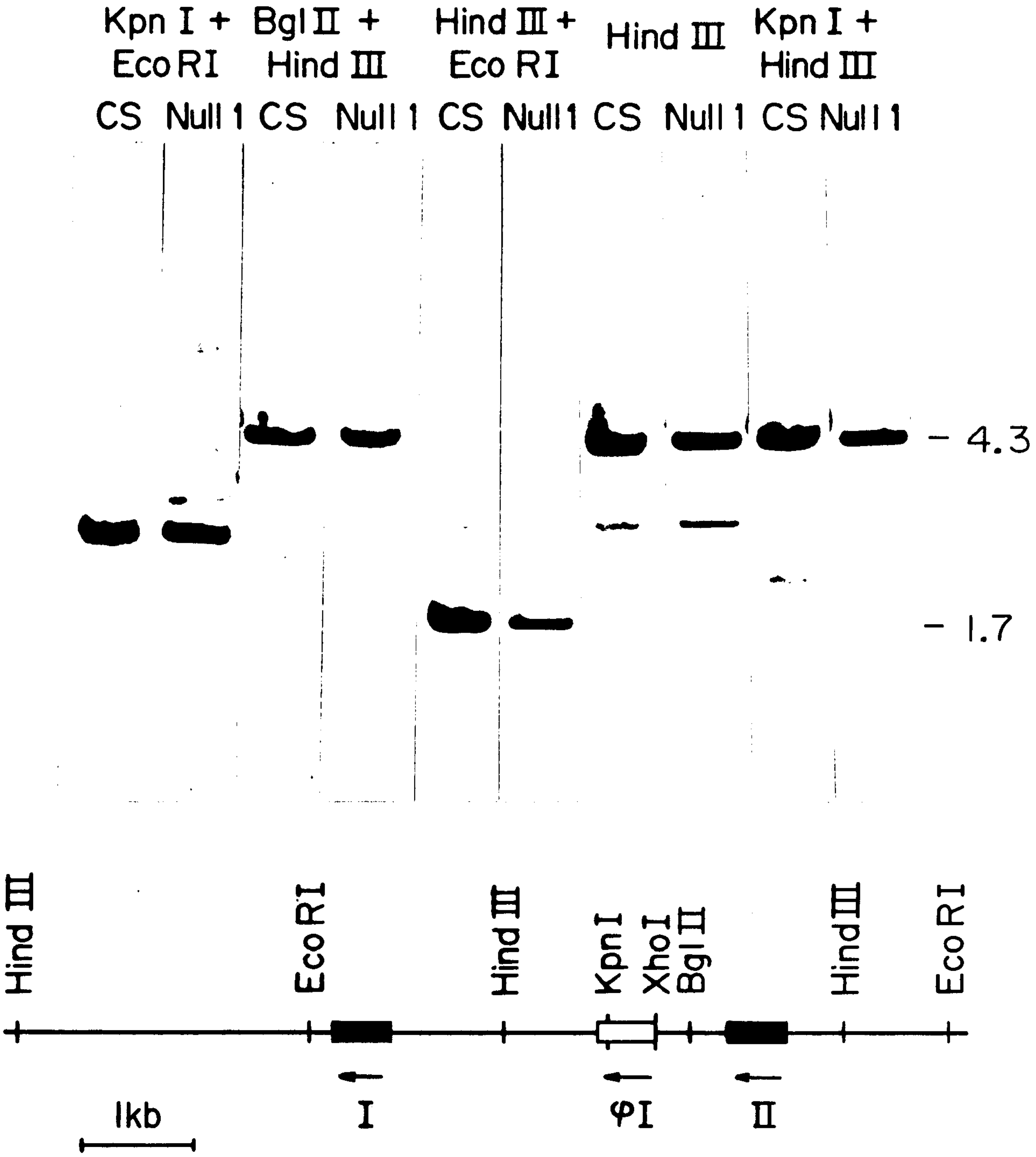

Kpn I + Eco RI
Bgl II + Hind III
Hind III + Eco RI
Hind III
Kpn I + Hind III
CS
Null 1
CS
Null 1
CS
Null 1
CS
Null 1
CS
Null 1
- 4.3
- 1.7
Hind III
Eco RI
Hind III
Kpn I
Xho I
Bgl II
Hind III
Eco RI
1kb
I
φI
II

Figure A7