
Sequence and expression of the *Cc* gene, a member of the dopa decarboxylase gene cluster of *Drosophila*: possible translational regulation

David D. Eveleth Jr and J. Lawrence Marsh*

Developmental Biology Center and Department of Developmental and Cell Biology, University of California, Irvine, CA 92717, USA

Received 13 January 1986; Revised and Accepted 27 June 1986

ABSTRACT

A transcript has been localized proximal to the dopa decarboxylase (*Ddc*) gene within a cluster of genes involved in cuticle formation and catecholamine metabolism in *Drosophila*. This gene, which has been identified as *l(2)37Cc*, maps 2.0kb from the 5' end of the *Ddc* gene and is transcribed in the same direction as *Ddc*. We describe a new deficiency which in conjunction with previous deficiencies localizes the *l(2)37Cb* and *l(2)37Cc* loci to the cytogenetic interval 5' to *Ddc*. We present the sequence of the *Cc* gene and corresponding cDNA. The *Cc* message contains several open reading frames 5' to the large open reading frame responsible for the lethal complementation group, suggesting that expression of *Cc* function may be regulated translationally. The *Cc* transcript is expressed in early embryos, late embryos, late third instar larvae and adults. We discuss the implications of these findings with respect to the gene organization in the region.

INTRODUCTION

Eukaryotic genes which serve related functions are often clustered in the genome (for example, globin genes (1); heat shock genes (2); chorion genes (3); and cuticle protein genes (4)). On the basis of genetic evidence it has been suggested that the region of the *Drosophila* genome surrounding the dopa decarboxylase (*Ddc*) gene contains a cluster of functionally related genes engaged in cuticle production and catecholamine metabolism. Of the 18 complementation groups mapped within the *Df(2L)TW130* region, 14 affect the formation of the cuticle in some way (5,6,7). In addition to *Ddc*, two other genes involved in catecholamine metabolism, di-phenol oxidase (*Dox-A2*) and *l(2)amd^H*, map in this region (6,7,8). This gene cluster is atypical since most gene clusters which have been described in metazoans encode structural proteins, e.g., cuticle proteins (4) or yolk proteins (9) rather than enzymes.

Approximately 100kb of DNA flanking *Ddc* has been cloned from the salivary gland region 37C (6). The DNA surrounding *Ddc* has been reported to be unique as analyzed by southern blot (6). Molecular analysis of deletion breakpoints indicates that the genes in this region are very densely clustered on the molecular scale (e.g., at least 6 genes within an 18-23kb region). Transcription mapping in the region 3' to the *Ddc* gene has assigned 3 genes to within 12kb which produce unspliced transcripts of at least 10.5kb(10). Although this physical organization has potential significance for regulation of these genes, the only genes which have been mapped at the molecular scale in this region are *Ddc*(11) and *l(2)amd*(12). To further elucidate the structure and organization of the genome in this region we have studied transcription from the region immediately adjacent to the proximal (5')end of the *Ddc* gene. In this report we describe the isolation of cDNA clones and mapping of these to the region immediately

Nucleic Acids Research

proximal to the 5' end of the Ddc gene. We present the complete sequence of this gene and a corresponding cDNA. We describe a new deficiency which localizes the l(2)37Cb and l(2)37Cc loci to the cytogenetic interval 5' to Ddc and we argue that this transcript corresponds to the l(2)37Cc locus. We show that transcripts from this locus are present throughout embryogenesis, at pupariation and in adults. These molecular mapping studies serve to further delimit the loci in this genetically well studied region.

MATERIALS AND METHODS

Reagents and enzymes:

Restriction enzymes, DNA polymerase and DNAase I were obtained from Boehringer-Mannheim or New England Biolabs. S1 nuclease was obtained from Bethesda Research Labs, nitrocellulose from Schleicher and Schuell, nylon membranes from AMF, oligo-dT cellulose from Collaborative Research, and labeled nucleotides (³²P) from New England Nuclear.

cDNA library screening:

Embryonic (1.5-5hr) and pupal (1 day pupae) cDNA libraries were the kind gift of Drs. D. Hogness and M. Goldschmidt-Clermont, Stanford University. Screening of libraries and recovery of phage was performed as described in (13). DNA was labeled by nick translation as recommended by NEN.

DNA Blotting:

DNAs were restricted and 1ug of DNA per lane was electrophoresed in 1.0% agarose in 100mM Trisborate buffer. Gels were denatured, neutralized and blotted onto nitrocellulose according to the method of Southern(14). DNA for genomic blots was extracted from flies by the method of Davis(15). DNA from 20 flies was restricted and used in one lane. Blotting onto nylon membrane was according to the manufacturer's recommendations.

Hybridization:

Nitrocellulose filters were hybridized in 4xSET with heparin(50ug/ml), SDS(0.1%), and sodium pyrophosphate(0.2%) (16). Filters were hybridized at the indicated temperatures for at least 24 hours. Washing was performed at the hybridization temperature in 0.5x SET. SET is 25mM NaCl, 1.5mM Tris pH=7.4, 0.1mM EDTA. Nylon filters were hybridized and washed according to the manufacturers instructions.

Sequencing:

DNA fragments to be sequenced were cloned into M13 vectors(17) and sequenced using the dideoxy chain termination procedure(18) using kits obtained from P-L Biochemicals.

Gene fusions:

Gene fusions were constructed by cloning appropriate restriction fragments into the polylinkers of the pUC(17) or pIC(19) vectors so as to create readthrough translation from the lacZ gene. The identity of each construct was confirmed by restriction analysis (data not shown). An example of each construct was grown overnight in B broth (17), diluted tenfold into B broth supplemented with 0.1mM isopropyl-b-D-thiogalactoside and grown for four hours. Cells from 10 ml of media were harvested by centrifugation and solubilized in 1 ml of 20% glycerol, 10% b-mercaptoethanol, 6% sodium dodecyl sulfate and 10mM

Tris pH=8.0. Debris was removed by centrifugation at 12,000g for 15 min and 25 ml of the supernatant was loaded onto a discontinuous polyacrylamide slab gel (20).

Northern blots:

Drosophila RNA was prepared from staged Oregon-R animals using the method of Geitz(21). Early (1-6hr) and late (16-24hr) embryos were collected from a population cage at 25 degrees. The late embryo collection contained approximately 10% hatched larvae. The late third instar collection was made by taking wandering larvae and white pre-pupae from the sides of uncrowded bottles. This sample contains less than 5% pupae. The adult animals were aged three to five days after collection. Poly-A⁺ RNA was selected by a single pass over oligo-dT cellulose. The RNA was ethanol precipitated and resuspended in 50% formamide, 2.2M formaldehyde, 10mM MOPS and heated to 55 degrees for 15 min. to denature. Poly-A⁺ RNA(15ug/lane) was electrophoresed in 1.0% agarose gels containing 2.2M formaldehyde. Denatured DNA was used as a size marker. The gels were blotted immediately after running onto nylon membranes using 20X SSC as the blotting buffer. Filters were probed with actin as a control for the transfer of intact RNA (data not shown).

RESULTS

Identification of the transcript:

We isolated cDNA clones from 1 gt10 cDNA libraries using a 7.2kb *Pst*I fragment of genomic DNA as a probe. The cDNAs were subcloned into pUC plasmids (17) and restriction mapped (Figure 1). The Cc1, Cc2 and Cc3 cDNAs were isolated from libraries constructed using early embryonic (1.5-5hr) poly-A⁺ RNA and the Cc4 cDNA was isolated from a library made with early pupal poly-A⁺ RNA. On the basis of the restriction data and the observation that one of the clones hybridizes to each of the others (data

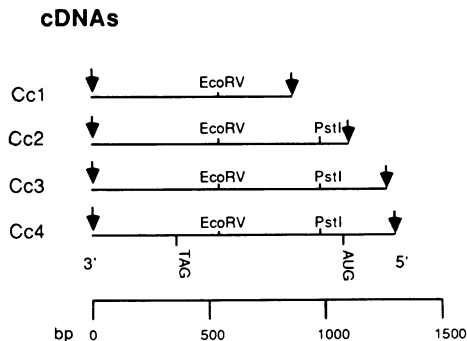


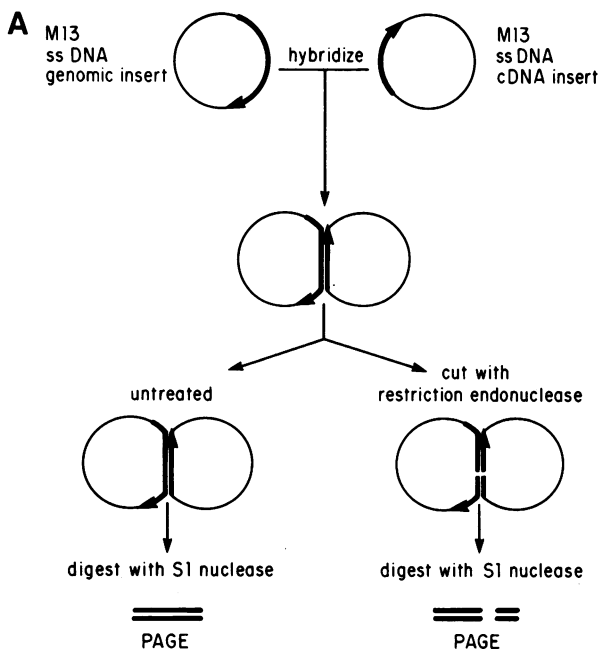
Fig. 1. Restriction maps of the cDNA clones. Arrows indicate the *Eco*RI linkers used in construction of the libraries. The cDNAs contain no sites for the enzymes *Hind*III, *Bam*HI, or *Cla*I. The positions of the initiator and terminator codons of the long ORF are shown. The Cc3 and Cc4 cDNAs both have 70-90 A residues at their 3' ends. In this and all subsequent figures the clones are shown oriented to the chromosomal DNA with the centromere to the right and transcription proceeding from right to left.

not shown), we conclude that these clones all represent the same transcript. No evidence for alternative processing variants was detected.

Mapping of the transcript onto the genome:

To locate the transcription unit represented by the cDNAs, we probed blots of cloned genomic DNA with nick-translated cDNAs at high stringency (data not shown). This analysis localized the region of transcription to the 1 to 5 kb interval proximal to *Ddc*. More precise mapping was accomplished using the S1 nuclease digestion strategy shown in Figure 2A. The Cc3 cDNA was subcloned into M13 and single-stranded phage DNA was isolated. This DNA was hybridized to single stranded M13 clones of genomic DNA from the same region in the opposite orientation (fragments A, B, and C in Figure 2). The resulting hybrids were digested with S1 nuclease and the protected fragments examined by polyacrylamide gel electrophoresis (PAGE). When the protected fragment lay completely within a genomic fragment, its end points were localized by cleaving the hybrid with a restriction endonuclease before S1 digestion (e.g. fragment A in Figure 2). Typical results of these experiments are shown in Figure 2B. This analysis identifies two exons in the Cc3 cDNA, one of about 400bp and one of about 800bp (Figure 2C).

The exact locations of the splice junctions and boundaries of the cDNA were determined by sequencing the Cc4 cDNA and comparing this to the genomic sequence. A diagram of the strategy used to generate the sequence is shown in Figure 3 and the sequence itself is shown in Figure 4. The majority of six base restriction sites predicted by the sequence have been tested and confirmed and the expression vectors constructed on the basis of the sequence (see below) behave exactly as predicted



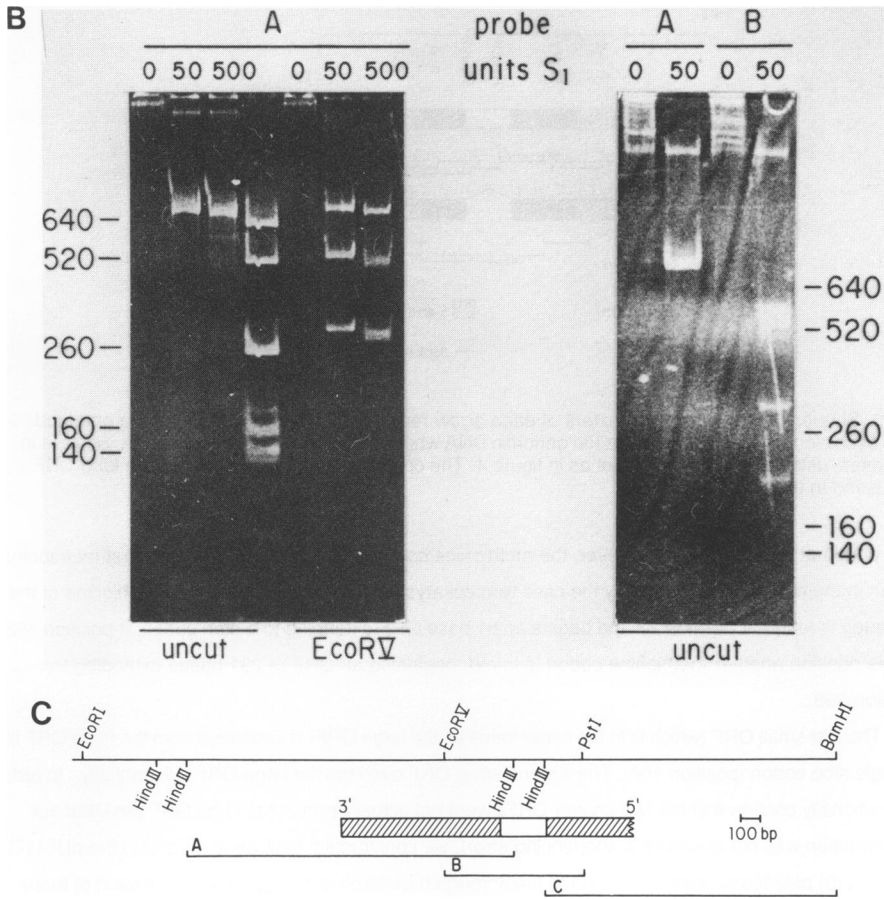


Fig. 2. Transcription mapping of the *Cc* gene. A) diagram of the method; B) results of S1 digestions; C) derived transcription map. The letters in C denote fragments used in part B. The experiment involving fragment c is not shown.

indicating that the sequence is highly accurate.

Comparison of the cDNA and genomic sequences identify two exons (Figure 3), one 789bp in length and one at least 380bp in length. Sequence analysis of the *Cc3* as well as the *Cc4* cDNAs reveal a poly-A tail of 70-80bp at the point indicated in Figure 4. This analysis orients the gene and shows that the *Cc* gene is transcribed in the same direction as *Ddc*. This is in complete agreement with the exon sizes determined by S1 mapping.

Analysis of the *Cc* gene product:

The sequence of the cDNA contains several open reading frames (ORFs). The longest of these begins at a methionine codon at position 201, extends across the splice junction and terminates at the

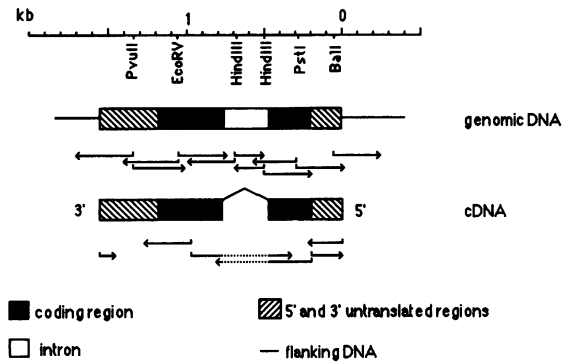


Fig. 3. Sequencing strategy. The extent of each arrow represents the extent of sequence obtained from each clone. The top diagram is the genomic DNA while the lower diagram is the cDNA. Scale is in kilobases, using the same zero point as in figure 4. The coding region shown is that of the long ORF discussed in the text.

stop codon at position 1173. However, the methionine codon at position 201 is not the first methionine codon in the message, as is usually the case with eukaryotic messages(22). The first methionine of the message is found at position 27 and begins an 81 base ORF extending to a stop codon at position 108. At this position another methionine codon is found, beginning an ORF of 291 bases extending to position 398.

The first small ORF (which is in the same frame as the large ORF) is separated from the large ORF by a single stop codon (position 108). The second small ORF overlaps the large ORF substantially. In order to functionally confirm that the two shorter ORFs were not actually part of the large ORF (and that our interpretation was not based on a sequencing error), we constructed lacZ gene fusions in the pUC(17) and pIC(19) plasmid vectors which lead to readthrough translation of the Cc message in each of three reading frames beginning at the BamHI site (position 53). An error of a single base in our sequence could lead to the translation of the large Cc ORF in these constructs, producing a protein of approximately 26kDa. Figure 5C+D show clearly that no new 26kDa protein is specifically encoded by one of these fusions. As a control for the fusion technique we also constructed similar fusions (Figure 5A+B) leading to readthrough transcription beginning at the PstI site (position 281), within the large ORF. One of these fusions specifically encodes a new protein of 24.3kDa, demonstrating that the large ORF is functional and confirming the sequence both for molecular weight of the translated protein and the reading frame at the PstI site. Each of these gene fusions utilized a fragment of the Cc4 cDNA extending to the 3' EcoRI site created by linker addition during construction of the cDNAs.

The Cc4 cDNA is full-length:

The Cc4 cDNA contains a long open reading frame which could be initiated at the AUG(methionine) codon at position 201. This AUG is the first AUG in this ORF and we conclude that this is the AUG initiator codon used to begin translation of the Cc gene product. Thus the Cc3 and Cc4 cDNAs span the entire coding region of the Cc gene. The Cc4 cDNA extends at least 184 bases upstream of the initiator

AUG. An additional approximately 20 bases at the 5' end of the Cc4 cDNA suggests that the 10 bases beginning at base 1 could have formed a foldback loop which may have served as the primer for second strand synthesis. Close examination of the genomic sequences just upstream from the 5' end of the cDNA reveal a consensus eukaryotic promoter including a TATA box (position -30, Figure 4) and a CAAT box (position -54, Figure 4). Thus the Cc4 cDNA appears to be full-length suggesting that the Cc gene initiates transcription at or very near the position marked as 1 in Figure 4.

The Cc gene is single-copy:

To determine if the Cc gene is repeated in the genome, we probed genomic Drosophila DNA with nick-translated Cc4 cDNA at high stringency (20 degrees below T_m, data not shown). This analysis showed that the only Cc complementary DNA in the genome is contained within a single fragment of the size predicted for the cloned region. In addition, in situ hybridization using probes containing the Cc locus identifies only one site of hybridization (23). These results clearly show that the Cc gene is single copy and that no closely homologous copies of Cc exist within the Drosophila genome.

Developmental expression:

To analyze the developmental regulation of expression of this new transcript, we probed RNA blots with nick translated Cc4 cDNA (Figure 6). Accumulation of the Cc transcript is greatest in the late embryo, somewhat less in early embryos, late third instar larvae and adults. Both Ddc and Cc transcripts are expressed during late embryogenesis, at pupariation and eclosion, but during early embryogenesis little Ddc mRNA is present (21,24,25). Thus the Cc transcript is regulated differently than Ddc. We find only a single approximately 1.6 kb species of Cc mRNA at all times of development. This is unlike Ddc in which substantial amounts of unprocessed message appear in the cytoplasm (21,24). We have seen no evidence for alternative processing of the transcript. If stage specific transcripts for this gene exist as they do for the adjacent Ddc (11) gene, the differences between transcripts must be small.

Cytogenetics:

Previous mapping (5) places the B gene series (i.e. (2)37Bx) distal and the C gene series (i.e. (2)37Cx) proximal to Ddc. A new deficiency has been isolated, Df(2L)hk^{UC2}, (Gibbs and Marsh, unpublished) which fails to complement hk function and is lethal over all the B genes as well as amd, Ddc, Cb and Cc. This deficiency in conjunction with previous deficiencies(6,7) and the mapping reported here establishes the following chromosomal order: [hk], [Ba.Bc.Be], [Bf.Bb], [Bg], [Bd.amd], [Ddc], [Cc], [Cb], [Ca.Cd.Ce.Cg], [Cf], (loci within brackets are unordered).

DISCUSSION

We have identified a transcript 5' to the Ddc gene of Drosophila by RNA blotting and by isolating and characterizing a series of cDNA clones. The longest cDNA (1.3kb) hybridizes to an approximately 1.6kb poly-A⁺ RNA species. The size of the longest cDNA is in agreement with the observed message size after addition of a poly-A tail. The 3' end of the transcript is specified by a poly-A tail present in the cDNA. Comparison of the orientation of the cDNA and genomic clones indicates that the Cc gene is transcribed in the same direction as Ddc (from centromere proximal to distal).

Nucleic Acids Research

TCGTGTTTG CATTTCAAAG AGCAAAATTG CGCAATAAATT TAATTTAAGT -200
 GCAACCAAAA
 ACCTTTAACA AGTGAGTCTA TGAGCGAGCC GGAGCTAACT -150
 GCGTTTGAAG GACCCAGGGA
 GAACAGGGGA TAAAAACACG CAAACAGCAC -100
 ATTGTCATTA AACCATGATC GGTTCTAACC
 TAAAAACTGA TGAAATTTCC AAGTCCAGTC -50
 GGCTTTCATT TACATATATA GTTTATAGAA
 CAGTAACTAA ATACATTACG TTTGTTTTAT >cDNA
 TCAAAGGATA **ATG**GCTGCTC AGTTCCTTAA
 >RNA begins **Met**AlaAla GlnPhePheAsn
 50 BalI 100
 TCGCATTGGC CAAATGGGCT CGGAGTGGCG TTTTGGGTGG CGTTGTCAAT TCGGCATTAT
 ArgIleGly GlnMetGly SerGluTrpArg PheGlyTrp ArgCysGln PheGlyIleIle
 150
 ATA**ATG**TGGA AGGCGGCCAC CGGGCGGTCA TCTTCGATCG CTTCACCGGC ATCAAGGAGA

MetTrp LysAlaAlaThr GlyArgSer SerSerIle AlaSerProAla SerArgArg
 200
 ACGTGGTCGG CGAGGGTACC CACTTCTTCA TCC**CA**TGGGT GCAGCGGCC ATCATCTTCG
 ThrTrpSer AlaArgValPro ThrSerSer **Met**Gly AlaAlaAla HisHisLeuArg
 SerHisGly CysSerGlyPro SerSerSer
 250 PstI
 GACCATCCGG TCCAGCCCC GCAACGTTC AGAGATAACG GGCAGCAAGG ATCTGCAGAA
 ThrIleArg SerGlnPro ArgAsnValPro GluIleThr GlySerLys AspLeuGlnAsn
 AspHisPro ValProAlaPro GlnArgSer ArgAspAsn GlyGlnGlnGly SerAlaGlu
 300
 TGTC AACATC ACGTCCGAA TCCTGTACCG CCCATTCCA GACCAGCTGC CCAAGATCTA
 ValAsnIle ThrLeuArg IleLeuTyrArg ProIlePro AspGlnLeu ProLysIleTyr
 CysGlnHis HisAlaProAsn ProValPro ProHisSer ArgProAlaAla GlnAspLeu
 350 400
 CACCATCTC GGCCAGGACT ACGACGAGCG TGTCTGCC TCCATCGCG CTGAGGTGCT
 ThrIleLeu GlyGlnAsp TyrAspGluArg ValLeuPro SerIleAla ProGlu|intron
 HisHisSer ArgProGlyLeu ArgArgAla CysProAla LeuHisArgAla ***
 450
 GAAGGCTGTG GTCGCCAGT TCGACGCCGG CGAGCTGATC ACCCAGCGTG AGGTGAGTAA
 500HindIII
 AAACATCAGG AGAGCCCTCA GTTCTTTAG GATAAAGCTT TTTTGTATGGG CCTTGTGTTT
 550
 CACCAGAGGC GCTTATGTTT ACTTAGTATC AGGTCATCAG ATCTCATCGT ATTCACGCG
 600
 TATTTTATTT TCGATTTAAA TTTGATTGTG TAAGCAAGAA CGCGATCTCG GCGCTTAAGT
 650 HindIII 700
 CTAGGCTTAT TGCTCTTTTG AGAATTGAAG CTTAAGAATT GGATTATTAT TCAACTTATG
 750
 CAGTCGGAAT TATAATACTT TCATGCCCGT ACTTAACGTG ATTACCTCTT TGCAGATGGT
 intron |MetVal

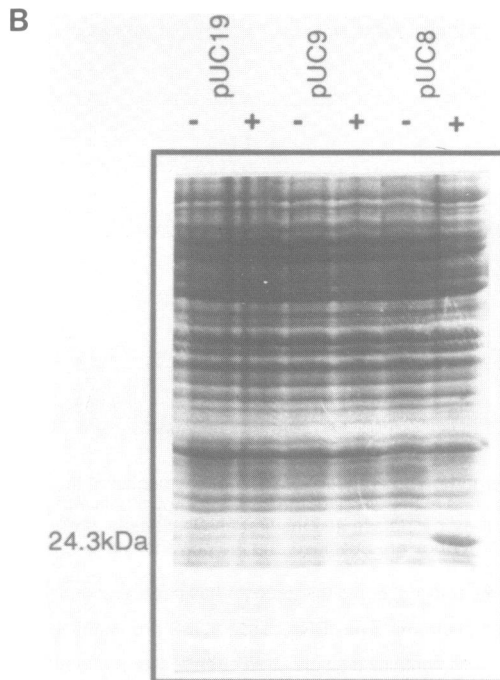
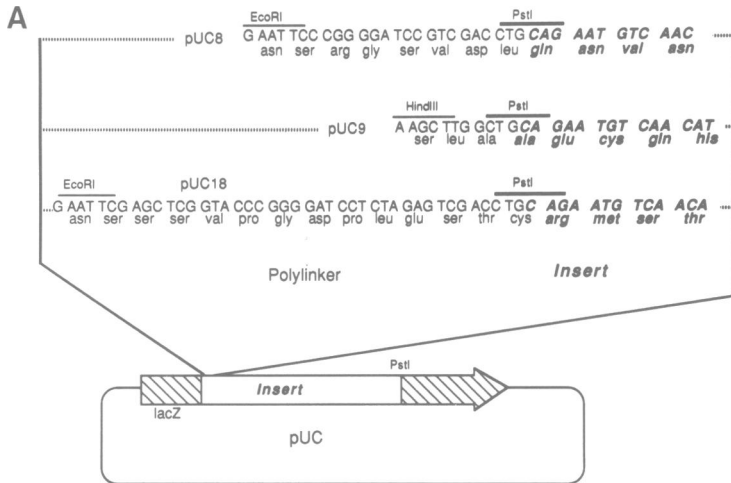
			800		
GTCGCAGCGC	GTTTCCCAGG	AACTGACTGT	ACGTGCCAAG	CAGTTCGGCT	TTATTCTGGA
SerGlnArg	ValSerGln	GluLeuThrVal	ArgAlaLys	GlnPheGly	PheIleLeuAsp
		850			
TGACATCTCG	CTCACGCACT	TGACCTTCGG	TCGGGAGTTC	ACGCTGGCCG	TCGAGATGAA
AspIleSer	LeuThrHis	LeuThrPheGly	ArgGluPhe	ThrLeuAla	ValGluMetLys
	900				
GCAGGTGGCC	CAGCAGGAGG	CGGAGAAGGC	GCGTTTTGTC	GTGGAGAAGG	CCGAGCAACA
GlnValAla	GlnGlnGlu	AlaGluLysAla	ArgPheVal	ValGluLys	AlaGluGlnGln
	950				1000
GAAGCTGGCG	TCCATTATTT	CGCGGAGGG	TGATGCCGAA	CGCGCTGTG	TTGGCCAAGT
LysLeuAla	SerIleIle	SerAlaGluGly	AspAlaGlu	ArgAlaCys	ValGlyGlnVal
				1050	<u>EcoRV</u>
CATTGCGAGG	CCGGAGACGG	TCTGGTGGAG	CCTGCGACTG	ATTGACCGGC	CGAGATATCG
IleAlaArg	ProGluThr	ValTrpTrpSer	LeuArgLeu	IleAspArg	ProArgTyrArg
			1100		
CCTCACCAGC	TATCCCCGGT	CCCGTGGAGT	CGCCTACTTG	CCCAGCGGAC	AGAGCCACGC
LeuThrSer	TyrProArg	SerArgGlyVal	AlaTyrLeu	ProSerGly	GlnSerHisAla
		1150			
TGCTCAATCT	GCCATCGACC	ATCGCGCAGT	AGCTGGGTGC	ATCTAGTTCC	GTTAAGTTGT
AlaGlnSer	AlaIleAsp	HisArgAlaVal	AlaGlyCys	Ile***	
	1200				
AACTAC2TAT	AGCATTTCAC	TAAGTACTTT	TCGATTTTGT	TTCTGCTGAA	ATATGCACTA
	1250				1300
CTCTAAAGCG	TTCGGCCGAC	TGACTGGAGA	ATACTAAGCG	AAACAACCAA	AATTTGTCTC
				1350	
ATGTAATCG	GTTTTTCCAT	TATCTTCCCG	ATCGGTTCGA	AATCCGGTCG	CAGTGTGGCC
			1400		
TCCGAATGCG	GCTCAACTTT	TGTTGTTTCC	GTTTGAGTGC	TTCGTGTGCG	CGTCCTGCTG
		1450			
AATTGGTCCT	TAAGATCGGC	ACCGGGCTTA	AAACA2CTCC	AGTAATTGAA	AGGAAAATTG
	1500				
AAAGGAAACC	ATAGACCGTA	AGAAAATTTC	ATTATGTGAA	CAAACAATAA	<u>AGTGATAATC</u>
	1550				1600
CCTACATCGA	GTGGGCTTTT	AAATTGAGTG	<u>TGTTATGGC</u>	TTATGCAGGG	AATGTTATTT
\polyA					
				1650	
AATTTATGAA	CACTTTTACT	TTCCA2CTTA	ATTACAGATC	GATCTAAAAG	CAAGTT

Fig. 4. Sequence of the *Cc* gene. The numbering of the bases begins at the postulated base #1 of the *Cc* RNA. Conceptual translation of the ORFs of the cDNA is shown below the sequence with the postulated initiator methionines in boldface and the first and last bases of the intron indicated by vertical lines. The TATA box and CAAT box are overlined and the polyadenylation signals are underlined.

Precise mapping of the *Cc* transcript onto the genome has been accomplished by a combination of S1 analysis and sequencing of cDNAs. The cDNAs encompass two exons separated by one intron (Figure 2). The cDNA sequence contains a single large open reading frame (ORF) beginning at 201 on

our sequence and extending to 1173 (Figure 4). Conceptual translation of this ORF predicts a protein of 24.8kDa. We have functionally confirmed the existence of this large ORF by constructing expression vectors producing the *Cc* gene product as a lacZ fusion protein (Figure5).

The sequence of the genomic DNA upstream from the 5' end of the cDNA is consistent with the



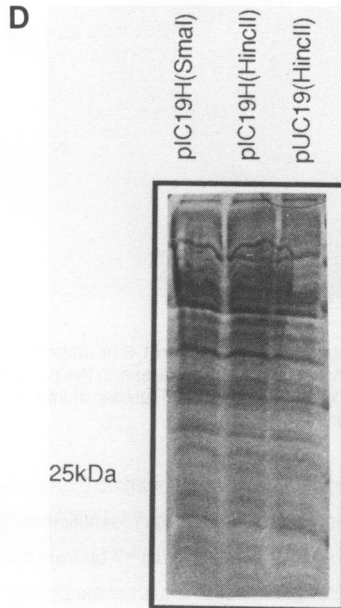
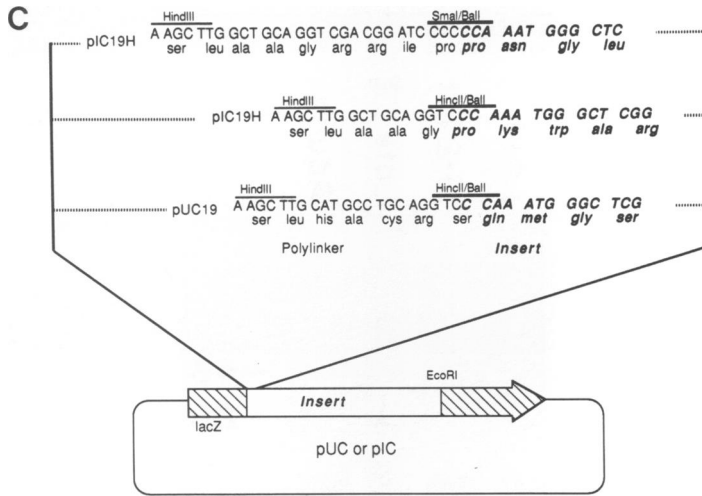


Fig. 5. Construction and analysis of *Cc* expression vectors. A) schematic of constructs testing the large ORF; the polylinker is on the left in plain text and the insert on the right in boldface italics; the *Pst*I site at the junction is indicated and the translation is shown below. B) Polyacrylamide gel analysis of proteins extracted from bacteria carrying these constructs, vector designated above; + denotes bacteria carrying the *Pst*I fragment in the orientation shown in part A; - denotes the inverted (antisense) orientation; the *Cc* fusion protein of 24.3kDa is indicated. C) schematic of constructs testing the existence of the stop codon of the premature ORF; polylinker and insert indicated as in A; the various sites used for fusion are indicated and the translation is shown below. D) polyacrylamide gel analysis of bacteria carrying the plasmids shown in C; the position of the protein expected if the stop codon of the first premature ORF does not exist is indicated by an arrow.

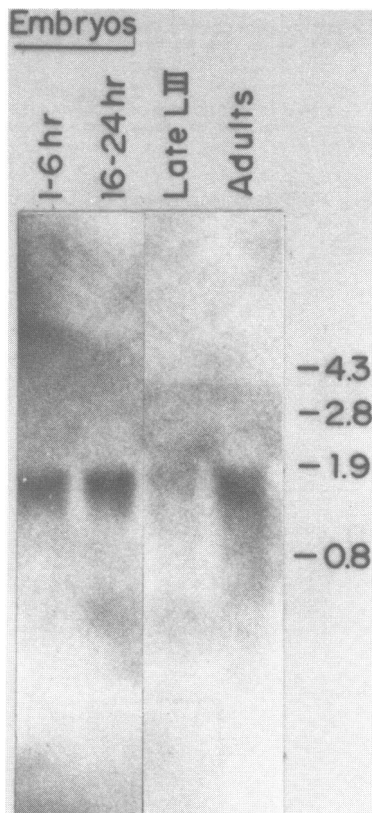


Fig. 6. RNA blot. Poly-A+ RNA (15ug/lane) isolated from 1-6 hr embryos, 16-24 hr embryos, late third instar larvae or adults was electrophoresed in 1.0% agarose in the presence of formaldehyde, blotted onto nylon and probed with nick-translated Cc4 cDNA. Transfer of intact RNA was confirmed by reprobing this blot with actin. Sizes in kilobases.

hypothesis that the Cc transcript begins at the position marked 0 on our sequence (Figure 4). This region contains a TATA homology and a CAAT homology (identified in Figure 4) in the appropriate positions to promote an RNA which would be initiated just 17 bp from the beginning of the cDNA. The Cc transcript is polyadenylated at a site 2.0kb from the start of the Ddc gene. The size of the cDNAs is in agreement with the observed message size after addition of a poly-A tail. This spacing of genes indicates that the primary transcripts of the Cc and Ddc genes do not overlap as is the case for the Ddc and Cs transcripts or the amd and Cs transcripts (25).

The Cc gene is expressed most strongly during late embryogenesis, with lower levels of transcript found in early embryos, late larvae and adults. *Drosophila* homozygous for lethal alleles of Cc die as larvae (T.R.F. Wright, personal communication). This indicates that the Cc gene product is dispensable for the development of properly formed larvae but that Cc is required for larval metabolism or for the

Summary of mapping in the 37C region

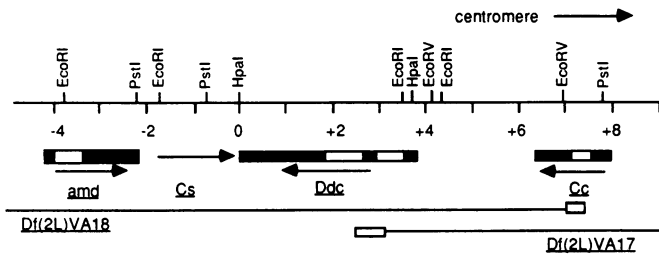


Fig. 7. Summary of mapping in the region. Position of the *Ddc* gene is from (11), and the *amd* gene is from (12). Solid areas represent regions incorporated into mature RNA; open boxes represent introns. Locations of the deletion breakpoints are from (6) with boxes indicating the uncertainty in the location of the deletion end point and lines indicating the region deleted. The line indicates the *Cs* transcript described in (22). The scale at top is in kilobases using the *HpaI* site at the 3' terminus of *Ddc* as an arbitrary zero point.

progression of the larva into a pupa. The finding that *Cc* is expressed in the late embryo but the lethal phase is at the larval stage is in contrast to *Ddc*, in which the peak of expression at late embryogenesis corresponds to the lethal phase.

The *Cc* gene is expressed at every time in development that *Ddc* is expressed as well as in the early embryo. The accumulation of *Cc* transcripts is low in late third instar larvae, a time when *Ddc* expression is maximal, and higher in aged adults, a stage where *Ddc* expression is restricted to the ovaries and central nervous system (26,27). The pattern of expression of the *Cc* gene is consistent with the hypothesis that *Cc* may function in cuticle synthesis since *Cc* is present at the molts and in early embryos, which may reflect carryover from ovarian expression. The physical organization which we have described indicates that these genes are so closely spaced that it is possible that putative enhancer-like elements of the *Ddc* gene may function to modulate the expression of the *Cc* gene or vice versa. The regulatory element of the *gypsy* transcription unit has been suggested to influence the *forked* locus over similar distances (28). Other regulatory elements are known to have domains large enough to function in this manner(29).

The occurrence of multiple methionine codons preceding the large major ORF is atypical. Functional analysis of our clones provide independent confirmation of the existence of these premature ORFs. Conceptually, any of the ORFs could correspond to the lethal complementation group *Cc*. Comparison of the *Cc* transcription map and the location of the breakpoint of deficiency *Df(2L)VA18* (6; Figure 7) indicates that the *VA18* deficiency deletes the 3' half of the *Cc* gene, almost certainly eliminating the function of the protein encoded in the large ORF but leaving the small, premature ORFs intact. The observation that some lethal alleles of *Cc* fail to complement *VA18* (and *VA17*; T.R.F. Wright, personal communication) shows that the large ORF is indeed responsible for the *Cc* complementation group.

Recent studies have shown that premature ORFs serve to regulate translation (30). The premature ORFs in the *Cc* message may also function to control translation although they are both much longer

than the ORFs observed to function in translational control in yeast. Translation of the first ORF would leave the ribosome exactly at the beginning of the second ORF, perhaps ensuring initiation of the second ORF. Translation of the second ORF would take the ribosome past the beginning of the large ORF. Thus the two premature ORFs may function to divert ribosomes from the large ORF. Alternatively, the structure of the Cc transcript may simply represent a system to produce several proteins from a single message.

The deficiencies reported here and the fact that some Cc alleles are not complemented by VA18 establishes Cc as the most centromere proximal gene in the VA18 deficiency (Figure 7). Further, P-mediated insertions of a genomic fragment containing the entire transcription unit described here rescue both lethal alleles of Cc and VA17/VA18 heterozygotes carrying transposed Ddc genes (J. Kullman and T.R.F. Wright, personal communication) thus confirming this conclusion.

ACKNOWLEDGMENTS

We would like to thank Ross Hodgetts, Charlotte Spencer, John Kullman, Ellen Pentz, and T.R.F. Wright for sharing their data before publication. We are indebted to M.Erfe for the original isolates of some of the cDNA clones and for her valuable and continuing contributions to this work and to Ross Hodgetts, Charlotte Spencer, Ken Konrad, P.D.L. Gibbs, T.R.F. Wright and John Kullman for permission to quote their unpublished observations. We would like to thank H.A. Schneidemann and the Monsanto Corporation for the facilities in which this work was performed. D. Eveleth is supported by PHS predoctoral training grant T32 CA09054. This work was supported by NIH grant GM28972 and PCM grant 8316485 to JLM.

*To whom correspondence should be addressed

REFERENCES

1. Efstratiatis, 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., Sholders, C.C., and Proudfoot, N.J. (1980). *Cell* 21:653-668
2. Corces, V., Holmgren, R., Freund, R., Morimoto, R., and Meselson, M. (1980) *Proc. Nat. Acad. Sci. USA* 77:5390-5393
3. Spradling, A.C., Digan, M.E., Mahowald, A.P., Scott, M., and Craig, E.A. (1980) *Cell* 19:905-914
4. Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J., and Davidson, N. (1982) *Cell* 29:1027-1040
5. Wright, T.R.F., Beerman, W., Marsh, J.L., Bishop, C.P., Steward, R., Black, B., Tomsett, A.D., and Wright, E.Y. (1981) *Chromosoma* 83:45-58
6. Gilbert, D., Hirsh, J., and Wright, T.R.F. (1984). *Genetics* 106:679-694
7. Pentz, E.S., and Wright, T.R.F. (1986). *Genetics* 112:843-859
8. Marsh, J.L., and Wright, T.R.F. (1980). *Dev. Biol.* 80:379-387
9. Riddell, D.C., Higgins, M.J., McMillan, B.J., and White, B.N. (1981). *Nucl. Acids Res.* 9:1323-1338
10. Eveleth, D.D. and Marsh, J.L. (1986). *Genetics*, in press.
11. Eveleth, D.D., Geitz, R.D., Spencer, C.A., Nargang, F., and Hodgetts, R.B., (1986). *EMBO J.*, submitted.
12. Marsh, J.L., Erfe, M.P. and Leeds, C.A. (1986). *Genetics*, in press.
13. Hunyh, T.V., Young, R.A., and Davis, R.W. (1984) in *DNA cloning techniques: a practical approach*, Glover, D., Ed., IRL press, Oxford.
14. Southern, E.M. (1975) *J. Mol. Biol.* 98:503-517
15. Davis, R.L. and Davidson, N. (1984) *Mol. Cell. Biol.* 4:358-367
16. Singh, L., and Jones, K.W. (1984) *Nucl. Acids Res.* 12:5627-5638
17. Messing, J. (1983). in *Methods in Enzymology*, Wu, R., Grossman, L., and Moldave, K., Eds., Acad. Press., New York, pp 20-78.
18. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). *Proc. Nat. Acad. Sci. USA* 74:5463-5467
19. Marsh, J.L., Erfe, M.P., and Wykes, E.J., (1984) *Gene* 32:481-485.
20. Laemmli, U.K., (1970) *Nature* 227:680-685

21. Geitz, R.D., and Hodgetts, R.B. (1985) *Dev. Biol.* 107:142-155
22. Kozak, M. (1981). *Nucl. Acids Res.* 9:5233-5252
23. Marsh, J.L., Gibbs, P.D.L., and Timmons, P.M. (1985). *Mol. Gen. Genet.* 198:393-403.
- 24: Beall, C.J., and Hirsh, J. (1984) *Mol. Cell. Biol.* 4:358-367
25. Spencer, C.A., Geitz, R.D., and Hodgetts, R.B. (1986). *Dev. Biol.* 114:260-264
26. Marsh, J.L., and Wright, T.R.F. (1979). in *Eukaryotic Gene Regulation*, Axel R., Manaitis, T., and Fox, C.F., Eds., Academic Press, New York, pp 183-194
27. Wright, T.R.F., Steward, R., Bentley, K.W., and Adler, P.N. (1981). *Dev. Genet.* 2:223-235.
28. Parkhurst, S.M., and Corces, V.G. (1985). *Cell* 41:429-437
29. Chandler, V.L., Maler, B.A., and Yamamoto, K. (1983). *Cell* 33:489-499
30. Fink, G.R. (1986). *Cell* 45:155-156