The EMBO Journal vol.6 no.9 pp.2803–2809, 1987

The transcription unit of the *Drosophila engrailed* locus: an unusually small portion of a 70 000 bp gene

Barry Drees, Zehra Ali, Walter C.Soeller, Kevin G.Coleman, Stephen J.Poole and Thomas Kornberg

Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA

Communicated by P.A.Lawrence

Transcripts from the *engrailed* gene of *Drosophila melanogaster* have been characterized by Northern, S1 nuclease sensitivity, and primer extension analyses. The *engrailed* gene encodes three poly(A)$^+$ transcripts (3.6 kb, 2.7 kb, and 1.4 kb) that derive from a 3.9-kb portion of the genome. No other transcribed regions were found up to 16 kb downstream and 48 kb upstream of the *engrailed* transcription unit, the portion of the genome to which *engrailed* mutations have been mapped. The structures of the *engrailed* transcripts are unaffected by lethal *engrailed* mutations that break the locus at points in the transcriptionally silent regions. Transcripts expressed by 145 kb of DNA that surround the *engrailed* locus were also identified by Northern analysis. In contrast to the large portion of the *engrailed* gene that is transcriptionally inactive, most of the surrounding regions are 10-fold more densely populated with transcripts. We presume that the unusually large silent region at the periphery of the *engrailed* transcription unit signify the presence of special mechanisms that regulate its expression.

Key words: Drosophila/engrailed/transcription

Introduction

An approach to a mechanistic understanding of developmental regulation and pattern formation has been to isolate mutations in key regulatory genes as a prelude to molecular analysis. An example is the *Drosophila engrailed* gene, in which mutations affect the processes that subdivide the fly into segments and compartments (Garcia-Bellido, 1975; Kornberg, 1981a). The *engrailed* gene function is required for the proper formation of the cellular blastoderm during the first hours of embryogenesis (Karr *et al.*, 1985), and for the development of the posterior but not the anterior compartment cells of each segment after cellularization (Lawrence and Morata, 1976; Kornberg, 1981b). Although the relationship between these pre-cellular and subsequent roles of the *engrailed* gene is as yet unclear, the partially 'anterior' phenotypes of posterior compartment cells that are mutant for *engrailed* suggest that, after cellularization, choice of an appropriate anterior or posterior developmental pathway depends upon *engrailed* function.

How genes such as *engrailed* control a developmental pathway is not known. Genetic analysis has shown that the gene is haplosufficient, that it provides an essential function to embryos and to later developmental stages, and that it presents a mutational target that approximates an 'average' *Drosophila* gene in size (Kornberg, 1981a). However, such descriptions do not indicate whether the affected *engrailed* gene function is a regulatory site where the products of other genes interact, or whether it is mediated through an RNA or a protein product. Our isolation of the *engrailed* gene in recombinant clones has made possible a molecular analysis of *engrailed* function. Fifteen chromosomal breakpoint *engrailed* mutations have been localized, and these mutations define a 70-kb region involved in *engrailed* function (Kuner *et al.*, 1985). The results presented here demonstrate the existence of only a single, small transcription unit within this large region.

Results

Expression of the engrailed gene

To determine which regions of the *engrailed* locus are transcriptionally active, the approximately 70-kb genomic region in which *engrailed* mutations have been localized (Figure 1) was subcloned

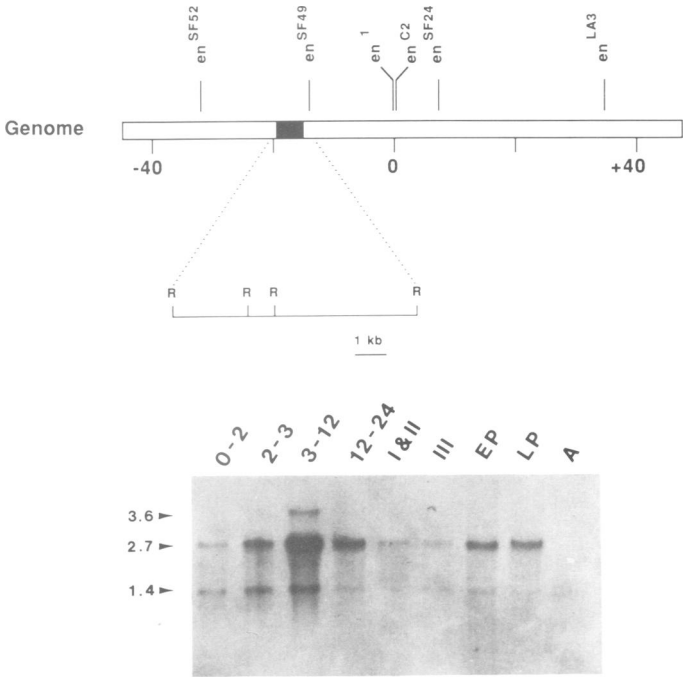

Fig. 1. The region of the *engrailed* locus is defined by the most proximal allele (en^{SF52}) and most distal allele (en^{LA3}) that have been localized in the genome (open horizontal box). The positions of the three *Eco*RI restriction fragments (2.9-kb, 0.9-kb and 4.7-kb) that contribute to the *engrailed* transcription unit (solid horizontal box) are indicated. For detection of transcripts, poly(A)$^+$ RNA was prepared from whole animals at various stages of development, fractionated on a formaldehyde–agarose gel (5 μg per lane), transferred to nitrocellulose, and hybridized with the 0.9-kb *Eco*RI genomic restriction fragment from within the coding region on the *engrailed* transcript. The developmental stages analyzed are as follows: (0–2), (2–3), (3–12), (12–24) h after egg-laying in the embryonic period; (I) first larval instar, 24–48 h after egg laying; (II), second larval instar, 48–72 h after egg laying; (III) third larval instar, 72–120 h; (EP), early pupal period, 120–168 h; (LP), late pupal period, 168–216 h; and (A) adults, 2–3 days post-eclosion. RNA sizes were estimated by comparisons with Rous sarcoma virus RNAs (2.9 kb, 4.8 kb and 9.5 kb) which were fractionated alongside the *Drosophila* RNA. Exposure was for 3 days.

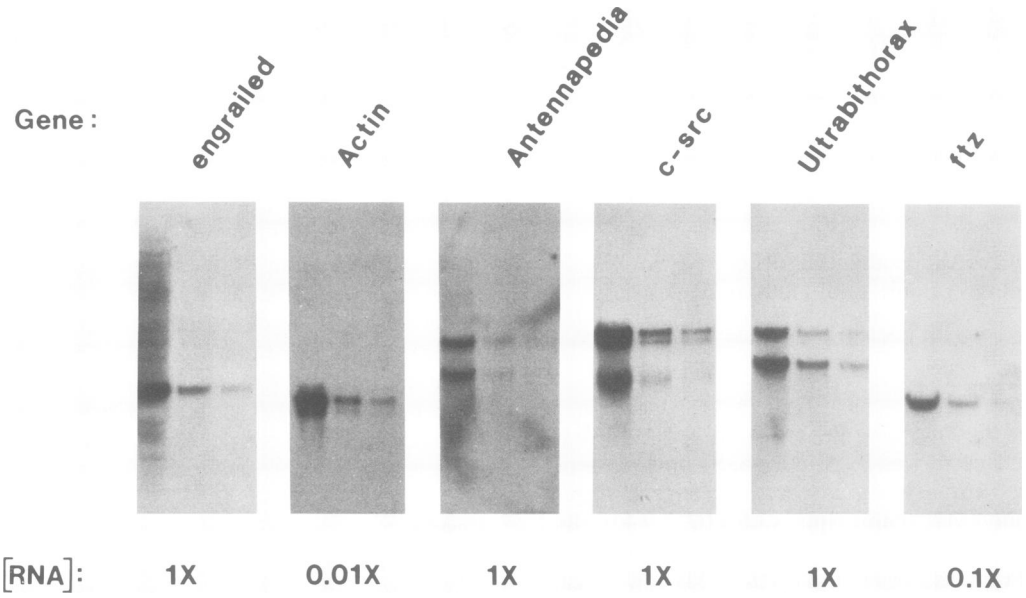

Fig. 2. Comparative abundance of the *engrailed* transcripts. Polyadenylated RNA was prepared from whole animals, fractionated, and transferred as in Figure 1. All comparative samples were from a single RNA preparation, with successive 5-fold dilutions. Radiolabeled probes were as follows: *engrailed* — a 2.0-kb *Eco*RI fragment from cDNA c2.4 containing 2000 bases from the 5′ end and including the homeo box; *Antennapedia* — a 2.3-kb insert from the λG1100 cDNA clone; *Ultrabithorax* — a 2.0-kb fragment from the pφ3602 cDNA clone; *fushi tarazu* (ftz) — a 1.8-kb insert from the pDmG2OR1.8 cDNA clone; *Drosophila* c-*src* — a 3.1-kb *Eco*RI fragment from the G4B-CA4 cDNA clone; actin — a 1.8-kb *Hin*dIII genomic fragment from λDmA2. Input counts for hybridizations were adjusted for the varying lengths of the restriction fragments, and all film exposures were identical (3 days).

into plasmid vectors as 23 *Eco*RI restriction fragments, and each subcloned fragment was purified and radiolabeled by nick-translation for use as probe. Each probe was separately hybridized to Northern blots of poly(A)$^+$ RNA isolated from embryos, larvae, pupae, and adults, and standardized for equivalent amounts of transcripts from a non-muscle actin gene (actin 5C, Tobin *et al.*, 1980; Fyberg *et al.*, 1980) to facilitate quantitative comparisons. RNA larger than 500 bp could be analyzed in these Northern blots. Probes prepared from three neighbouring *Eco*RI restriction fragments (2.9 kb, 0.9 kb, and 4.7 kb; Figure 1) from the centromere-proximal portion of the locus detected homologous RNA sequences. No transcripts were detected by any of the other 20 restriction fragment probes. The probes from the 0.9-kb and 4.7-kb fragments detected transcripts of 3.6 kb, 2.7 kb, and 1.4 kb in length; probe from the 2.9-kb fragment detected only the 3.6-kb and 2.7-kb transcripts. The three *engrailed* poly(A)$^+$ RNAs were detected in RNA isolated from cellular blastoderm stage embryos (2−3 h after oviposition), at high levels in gastrulation stage RNA (3−12 h), and at low levels in RNA from later embryos, larvae, and pupae. Transcripts were not detected in RNA prepared from adults (Figure 1). This developmental profile of expression corresponds with the period of action of *engrailed* function suggested by developmental studies of *engrailed* mutants (Lawrence and Morata, 1976; Kornberg, 1981a,b).

In order to estimate both the sensitivity of these measurements and the relative abundance of *engrailed* transcripts, the level of the 2.7-kb *engrailed* RNA was compared with the levels of transcripts of other *Drosophila* genes (Figure 2). RNA was monitored from the genes of *engrailed*, non-muscle actin 5C, the *Drosophila* c-*src* homolog, *Antennapedia*, *Ultrabithorax*, and *fushi tarazu* (*ftz*). For each determination, DNA probes were synthesized from restriction fragments approximately 2.0 kb in size. Developmental profiles of expression revealed that transcripts from all but the *ftz* gene were present at or near maximum levels in RNA isolated from embryos 3−12 h post-oviposition; *ftz* expression was maximal 2−3 h after egg laying. For direct comparisons, 1.5×10^7 c.p.m. of nick-translated probe from each gene were hybridized to RNA blots containing varying amounts of RNA from embryos of the developmental period of maximal expression. The *engrailed*, c-*src*, *Antp*, and *Ubx* genes yielded approximately equivalent hybridization signals, whereas actin transcripts could be detected in lanes with 1−2% as much RNA and *ftz* in lanes with 10−20% as much RNA as the other genes. The *engrailed* transcripts, as well as those of c-*src*, *Antp*, and *Ubx* are thus 50−100 fold and 5−10 fold less abundant than actin and *ftz*, respectively. If the *Ubx* transcripts represent 0.1−0.01% of the total embryonic poly(A)$^+$ RNA (Akam, 1983), then the 2.7-kb *engrailed* transcript represents a similar fraction of the total poly(A)$^+$ RNA and the minor *engrailed* transcripts approximately an order of magnitude less. Portions of the *engrailed* locus that appeared to be transcriptionally inactive apparently do not produce RNA larger than 500 bp in amounts greater than $1/10^5$ of poly(A)$^+$ RNA extracted from whole embryos.

To establish the polarity of the three *engrailed* RNAs, single-stranded probes were synthesized from a 232-nucleotide restriction fragment of the cDNA c-2.4 (nucleotides 1473−1715; Poole *et al.*, 1985) that had been subcloned in both orientations in the single-stranded phage vector M13mp9 (Messing and Viera, 1982). Probes for each strand were separately hybridized to duplicate Northern blots containing poly(A)$^+$ RNA from gastrulating embryos. The 3.6-kb, 2.7-kb and 1.4-kb transcripts are synthesized in the (centromere) distal to proximal direction. No transcripts were detected from the opposite direction. This result is consistent with sequence analysis of genomic DNA and cDNA clones of the 2.7-kb RNA (Poole *et al.*, 1985).

The 3.6-kb, 2.7-kb, and 1.4-kb transcripts differed in abundance and structure. The 2.7-kb transcript was at an approximately 10-fold higher concentration than either the 3.6-kb or 1.4-kb RNAs, and their relative abundance did not change during development. However, the precise structural differences bet-

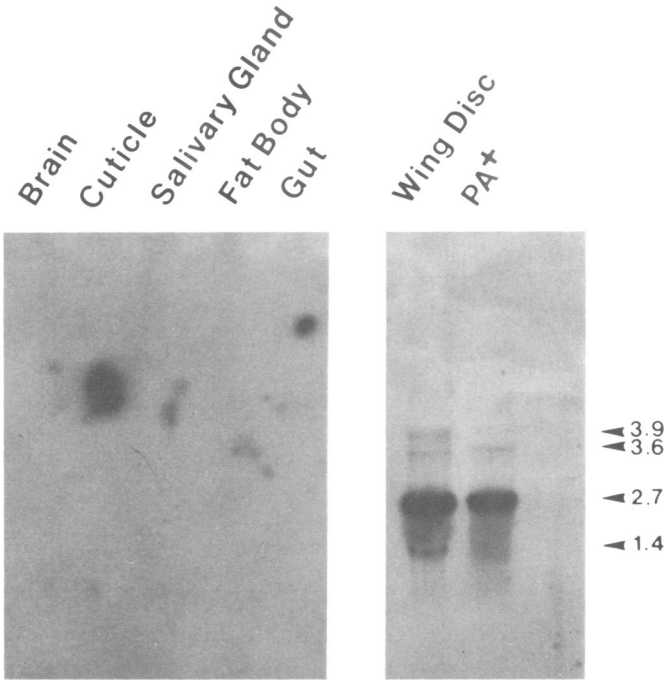

Fig. 3. Tissue specific expression of the *engrailed* transcript. Total RNA was isolated from the relevant tissues of 50 third instar larvae, fractionated on a formaldehyde–agarose gel, transferred to MSI nylon, and hybridized with a 230 b single-stranded probe which detected all the *engrailed* transcripts. Tissues were dissected by hand from wild-type Oregon R larvae. Tissues represented are: (PA^+) — 10 ng polyadenylated RNA from 3–12 h embryos; (Wing Disc) — wing imaginal discs; (Gut) — midintestine and proventriculus; (Fat) — fat bodies; (Sal. Glands) — salivary glands; (Cuticle) — cuticle stripped of all discs and internal organs, probably containing abdominal histoblasts; (Brain) — brains, including optic lobes and cephalic ganglion. Exposure was for 14 days.

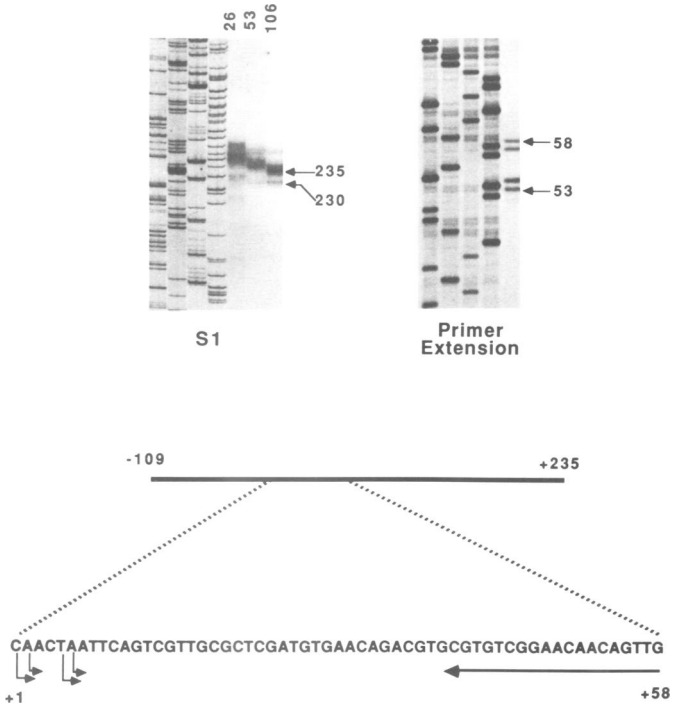

Fig. 4. S1 Nuclease and primer extension mapping of the 5′ end of the *engrailed* transcript. S1 nuclease analysis reveals a discontinuity at or near the presumptive start site of transcription (above left). A 344-nucleotide genomic fragment (nucleotides +235 to −109 indicated by solid line, below) uniformly labeled and single-stranded, was hybridized with 0–12 h poly(A)$^+$ RNA. The hybrid molecules were digested with either 26, 53 or 106 U of S1 nuclease. Increasing amounts of enzyme reduced the size of the protected fragments to ~235 and 230 nucleotides. An unrelated sequencing ladder was used for calibration. Primer extension analysis identifies the start of *engrailed* transcription (above right). An end-labeled synthetic oligonucleotide (nucleotides +58 to +39 indicated by horizontal arrow, below), downstream from the presumptive start site, was hybridized to 0–12 h poly(A)$^+$ RNA and extended with reverse transcriptase. The sizes of the extended products were calibrated with a sequencing ladder using the same primer, and indicate start sites at +1, +2, +5 and +6 (small arrows, below).

ween the 2.7-kb RNA and the two minor transcripts have not been determined. The most proximal genomic restriction fragment probe (2.9 kb) detected only the 3.6-kb and the 2.7-kb transcripts, suggesting that the 1.4-kb RNA has a different 3′ end. A probe synthesized from the smaller 3′ intron that is excized from the 2.7-kb RNA hybridized to only the 3.6-kb and 1.4-kb RNAs (not shown). Therefore alternate splicing pathways may produce the different transcripts. cDNAs representing the rarer transcripts were not found in our libraries.

In situ hybridizations to tissue sections of embryos and larvae and to isolated whole imaginal discs have demonstrated that *engrailed* expression is limited primarily to the cells of the posterior developmental compartments (Kornberg *et al.*, 1985). However, the probe used for the hybridization reactions detected all three transcripts, and to determine whether the individual species of *engrailed* RNA are produced in a cell-type specific manner, RNA was isolated from tissues of third instar larvae and subjected to Northern analysis. Although the overall amount of all three transcripts was observed to vary in different tissues, the relative concentrations of the three RNAs did not. High levels of the transcripts were detected in imaginal discs, but little or none in brain, cuticle, salivary glands, gut, or fat bodies (Figure 3). This result is consistent with the observation that, after *in situ* hybridization to sections of third instar larvae, *engrailed* transcripts could be detected only in the imaginal discs (Kornberg et al., 1985), and demonstrates that the predominant RNA characterized by *in situ* hybridization is the 2.7-kb transcript.

Embryonic RNA was tested for the presence of non-polyadenylated transcripts originating from the *engrailed* region. Probes from all 23 *Eco*RI restriction fragments from the *engrailed* locus were hybridized to blots of total RNA. In addition to the three poly(A)$^+$ *engrailed* transcripts, a 3.9-kb RNA was detected by the 4.7-kb, 0.9-kb, and 2.9-kb restriction fragments. No additional transcripts were detected with any of the 20 other probes (data not shown; the 3.9-kb RNA is visible in Figures 3 and 5). A primary *engrailed* transcript of approximately 3.9 kb is consistent with: (i) the production of a 2.7-kb mature transcript after excision of two introns, 1.1 kb and 0.28 kb in length (Poole *et al.*, 1985); (ii) sequence analysis of cDNA clones of the 2.7-kb transcript and the corresponding genomic regions (Poole *et al.*, 1985); and (iii) S1 nuclease sensitivity and primer extension analysis to identify the 5′ end of the 2.7-kb transcript (see below).

The initiation site for the 2.7-kb engrailed transcript

To localize the 5′ end of the transcription unit that generates the 2.7-kb *engrailed* RNA, poly(A)$^+$ RNA isolated from 3–12 h embryos was annealed to a uniformly labeled DNA fragment that spans the presumptive start site (nucleotides −109 to +235; see Figure 4). The hybrid molecules were digested with several different concentrations of S1 nuclease and fractionated electrophoretically. The sizes of the protected fragments (235 and 230 nucleotides; Figure 4) indicate either a transcription start or splice site approximately 213 bp upstream of the first ATG of the *engrailed* open reading frame (Poole *et al.*, 1985).

To determine if the 5′ end of the *engrailed* transcription units was located at this site, an end-labelled synthetic oligonucleotide complementary to nucleotides 12–26 of the cDNA c-2.4 (Poole *et al.*, 1985) was annealed to 3–12 h embryonic poly(A)$^+$ RNA and extended with reverse transcriptase. Electrophoretic fractionation indicated that the primer had been extended in almost equivalent amounts to positions 213, 212, 210 and 209 nucleotides upstream of the translation initiation codon (Figure 4). Therefore, the initiation site for the most abundant *engrailed* transcript apparently consists of a cluster of four positions, the most distal of which has been designated as position 1 (Figure 4).

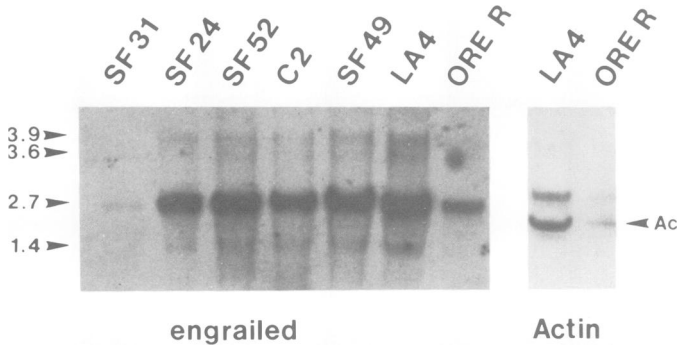

Fig. 5. Transcript analysis of *engrailed* mutants. RNA from 50 embryos was fractionated on a formaldehyde–agarose gel, transferred to MSI nylon, and hybridized with a 230-bp single-stranded probe which detected all the *engrailed* transcripts. With the exception of the homozygous deficiency (*Df (2R) en*^{SF31}) embryos (SF31), all of the embryos were *Df (2R) en*^{SF31} hemizygotes. The mutants analyzed were: (LA4) — *en*^{LA4}, a putative point mutant; (c2) — *In(2R)en*^{C2}, a chromosomal breakpoint mutation 14 kb upstream of the *engrailed* transcripts; (SF49) — *In(2R)en*^{SF49}, a chromosomal breakpoint mutant 2–5 kb upstream of the *engrailed* transcripts; (SF52) — *T(2,3)en*^{SF52}, a chromosomal breakpoint mutation 12–14 kb downstream of the *engrailed* transcripts; (SF24) — *T(2,3)en*^{SF24}, a chromosomal breakpoint mutation ~20 kb upstream of the *engrailed* transcripts; (+) — *CyO*, the second chromosome balancer, CyO. RNA samples were also monitored for non-muscle actin transcripts, and all of the lanes which contained RNA extracted from wild-type Oregon R embryos. Exposure for the *engrailed* probe was for 10 days.

Transcript expression in engrailed mutants

To investigate the molecular basis of the phenotypes of *engrailed* mutations, RNA from mutant embryos was analyzed with single-stranded probes homologous to *engrailed* transcripts. Mutant embryos were produced by crossing heterozygous *engrailed* strains to strains heterozygous for an *engrailed* deletion, and *en*/Df(*en*) embryos from the cross were selected under a dissecting microscope by their unusual and asymmetric gastrulation (Karr *et al.*, 1985). Three types of mutations were examined: a deficiency for the entire *engrailed* region (*Df(2R)en*^{SF31}), putative point alleles (*en*^{LA4} and *en*^{LA7}), and breakpoint alleles (*In(2R)en*^{C2}, *T(2;3)en*^{SF24}, *In(2R)en*^{SF49}, and *T(2;3)en*^{SF52}). A portion of the selected mutant embryos of each genotype was allowed to develop for 24 h; greater than 90% confirmed their mutant designation. RNA was extracted from the remaining portion of mutant embryos for analysis. RNA preparations from homozygous deletion embryos had levels of *engrailed* transcripts that were less than 10% wild-type. We attribute this low level to the small fraction of the selected embryos that were incorrectly identified. In contrast, the *engrailed* transcripts (3.6 kb, 2.7 kb, 1.4 kb and 3.9 kb) from the putative point and break-

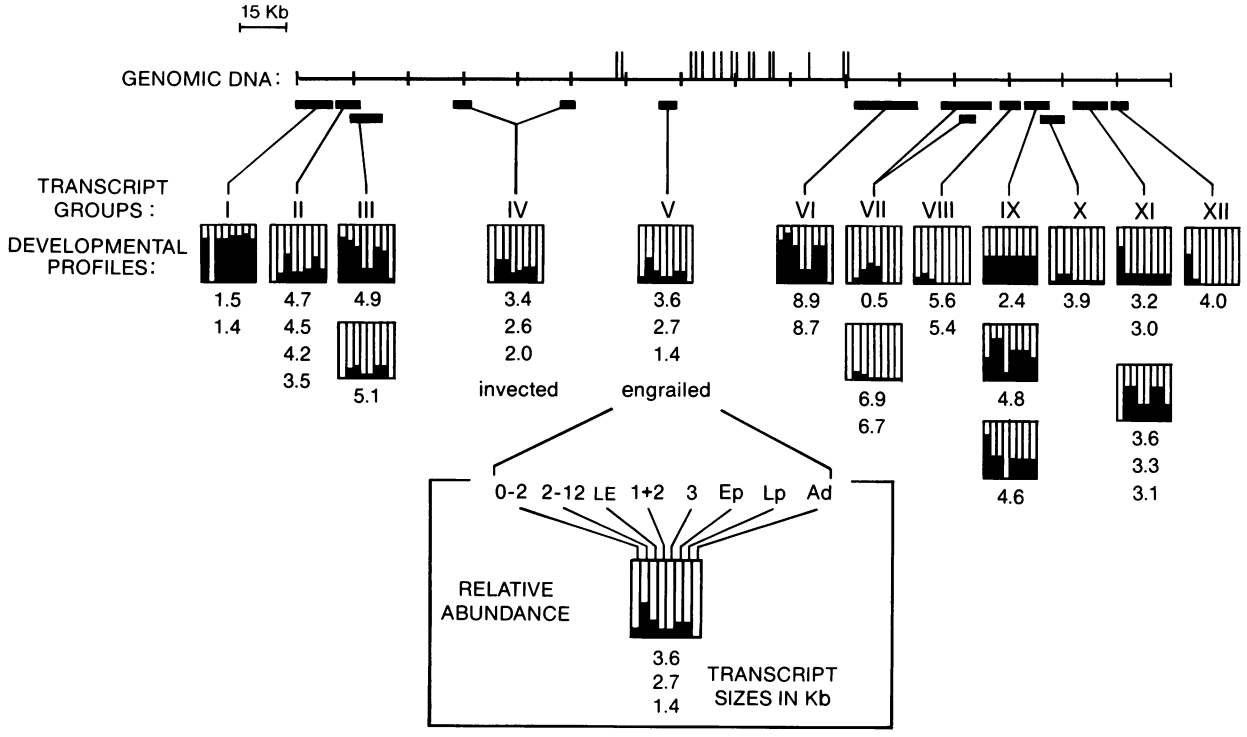

Fig. 6. Transcriptional profile of the *engrailed* complex and surrounding regions. The top horizontal line represents the 215 kb of genomic DNA analyzed in this study. The vertical lines above the genomic DNA indicate the location of *engrailed* breakpoint mutations. Horizontal bars below the genomic DNA represent the location of the DNA fragments which detected the RNAs of each transcript group. Each transcript group is designated by roman numerals and the developmental profiles of the transcripts are displayed in the bar graphs below. The transcript lengths are listed beneath each bar graph in kb, and the relative abundance of the transcripts is compared to the cellular actin gene (maximum height), the *engrailed* 2.7-kb transcript at 2–12 h (one-half maximum height), and the transcript group VII:6.9 transcript at 12–24 h (minimum height). Developmental times are as listed in the enlarged bar graph and are similar to those in Figure 1, except that 12–24 h is listed as LE for the late embryo.

point mutant embryos were not detectably altered (Figure 5). We conclude that there are no small exons outside of the proposed *engrailed* transcription unit, and that the breakpoint mutations affect the function of the *engrailed* gene without disturbing the structure of its transcripts.

Expression of the regions surrounding the engrailed gene

To better evaluate the significance of the structure of the *engrailed* gene, with its unusually large non-transcribed regions 5′ and 3′ of the transcription unit, the expression of the chromosomal regions including and surrounding the *engrailed* locus were compared. This portion of the *Drosophila* genome includes the *engrailed* and the engrailed-related *invected* genes (Poole *et al.*, 1985; Coleman *et al.*, 1987), and several other lethal complementation groups (Ali,Z. and Kornberg,T. unpublished results). All 61 of the *Eco*RI restriction fragments representing the 215 kb of DNA that had been cloned from the 48AB region of the genome were individually subcloned into plasmid vectors. Nick-translated probes were prepared from the gel-purified restriction fragments excised from these plasmids, and were used to monitor the presence of poly(A)$^+$ transcripts in RNA preparations from embryos, larvae, pupae and adults.

Although approximately one-half (31 of 61) of the restriction fragment probes hybridized to homologous RNA sequences, the distribution of the transcribed regions in the genome was non-random (Figure 6). Most of the transcripts originated from the regions flanking *engrailed* and *invected*. As shown above, only 3 of 20 *engrailed* fragments, encompassing 72 kb, detected transcripts, and 2 of 11 *invected* fragments, encompassing 35 kb, detected transcripts from this *engrailed*-related gene (Figure 6; Coleman *et al.*, 1987). In contrast, 26 of 30 restriction fragments that are either proximal or distal to these genes contain genomic sequences that are transcriptionally active. These restriction fragments detected 10 different groups of transcripts in the 108 kb of DNA that surrounds *engrailed* and *invected*.

Thirty-one transcripts that differed in size and temporal pattern of expression were detected. These transcripts are not spaced singly along the chromosome, but appear to overlap, since many individual restriction fragments hybridized to multiple RNA species. Only two regions were found with no transcription length heterogeneity. The source of transcript length heterogeneity has not been identified, nor has the precise number of transcription units. All but one of the transcripts detected varied in concentration during development. These results are summarized in Figure 6, and a brief description of each transcript group follows in the order they were detected by probes from successively more distal genomic sequences:

(i) Two transcripts (1.5 kb and 1.4 kb) were relatively abundant in all periods except the mid-embryonic, from 2−12 h. Their apparent presence prior to cellular blastoderm formation (at which time a rapid increase in embryonic transcription occurs) and absence afterward, suggests that they constitute part of the maternal contribution to the embyro.

(ii) Four transcripts (4.7 kb, 4.5 kb, 4.2 kb and 3.5 kb) were present from 2−12 h post fertilization to the adult stage in varying concentrations. Their greatest abundance was during the late embryonic and pupal stages.

(iii) Two transcripts (5.1 kb and 4.9 kb) were present throughout the life cycle, with greatest abundance during the embryonic and pupal periods.

(iv) Three transcripts (3.4 kb, 2.6 kb and 2.0 kb) of the *invected* gene (Coleman *et al.*, 1987) were present in varying concentrations throughout the life cycle except in very young, pre-blastoderm embryos and in adults. Their developmental profile is similar to that of the *engrailed* gene, except that the period of abundant expression extended later into the embryonic period.

(v) The *engrailed* gene transcripts (see above).

(vi) Two large and abundant transcripts (8.9 kb and 8.7 kb) varied considerably in concentration during the embryonic, larval, and pupal stages. They are adjacent to, but do not overlap, the most distal *engrailed* breakpoint mutations.

(vii) Two transcripts (6.9 kb and 6.7 kb) that are homologous to several restriction fragments spanning 14 kb overlap with a 0.5 kb transcript that can be localized only to an internal 1.4 kb region. The 6.9 kb and 6.7-kb transcripts were present in all post-blastoderm stages; the 0.5-kb RNA was present in greater abundance, but only during the larval and post-blastoderm embryonic stages.

(viii) Two transcripts (5.6 kb and 5.4 kb) were present during all embryonic but not subsequent stages.

(ix) Three transcripts (4.8 kb, 4.6 kb and 2.4 kb) were present during all developmental stages. The 2.4-kb RNA did not appear to vary in concentration; this was the only RNA species detected that was without apparent developmental regulation.

(x) One transcript (3.9 kb) was present at slightly elevated levels during embryonic stages and was at low concentration throughout the remainder of the lifecycle.

(xi) Two transcripts (3.2 kb and 3.0 kb) were most abundant in very young embryos and throughout the subsequent lifecycle. Three transcripts (3.6 kb, 3.3 kb and 3.1 kb) were not present in very young embryos but appeared subsequently; they were most abundant during the later embryonic and pupal stages.

(xii) One transcript (4.0 kb) was present only in young embryos.

Discussion

These studies have established that the *engrailed* locus, originally identified by mutant phenotypes, codes for a small family of poly(A)$^+$ RNAs. Several observations indicate that these RNAs embody the *engrailed* function.

Temporal and spatial programs of *engrailed* transcript synthesis are consistent with the demonstrated requirement for *engrailed* function during development. *engrailed* mutants develop abnormally from the first hours of embryogenesis, and studies of genetically mosaic flies indicate a continuing requirement for *engrailed* function until the adult stage. This requirement is limited to the cells of the posterior developmental compartments. The three *engrailed* poly(A)$^+$ transcripts were found in embryos, larvae and pupae, but not in adults. The transcripts were detected in imaginal discs, but not in third larval instar central nervous systems, salivary glands, fat bodies, cuticle, digestive tract, or adult ovaries. Thus, only those cells known to require *engrailed* function for their normal development have been shown to express the *engrailed* gene, and they do so during the relevant developmental periods. Regarding the organs in which no expression was observed, we can conclude only that *engrailed* expression must be less than the level of detection in our analysis.

Although the *engrailed* RNAs have not been shown to be polysomal, sequence analysis of a cDNA copy of the 2.7-kb *engrailed* transcript reveals an open reading frame that can potentially encode a protein of ~60 kd (Poole *et al.*, 1985). Furthermore, antibodies raised against a hybrid protein composed of *Escherichia coli* β-galactosidase and this *engrailed* open reading frame bind to a protein of ~60 kd produced by *Drosophila* embryos (T.Karr, N.Gay and T.Kornberg, unpublished) and to antigen present in embryonic posterior compartments (DiNardo *et*

al., 1985). We conclude that the *engrailed* function is likely to be a product of at least the 2.7-kb poly(A)$^+$ RNA, the most abundant transcript identified in this work.

The functions of the two other RNAs produced by the *engrailed* transcription unit are not known. Throughout development, the 1.4-kb and 3.6-kb transcripts were present in the same relative proportions to the 2.7-kb RNA; in RNA preparations from whole animals, their abundance is very low. They may be aberrant splicing products of the primary transcript.

The most unusual feature of the *engrailed* locus is the size of the transcription unit relative to the chromosomal interval in which *engrailed* mutations are located. *engrailed* mutations that physically break the locus (e.g. inversions or translocations) are distributed over almost 70 kb. Yet, Northern analysis of this 70-kb region detected homologous RNA exclusively from an interval of only 4 kb. Evidence for the size of the transcription unit and the absence of more proximal or distal small exons is as follows: (i) breakpoint mutations that physically separate the locus into two parts do not alter the size of the transcripts; (ii) comparative analysis of genomic sequences and apparently full length cDNAs localize the 5′ and 3′ ends of the 2.7-kb transcript within the designated transcription unit; (iii) primer extension and S1 nuclease analyses localize the 5′ ends of the 2.7-kb transcript close to those defined by the cDNA clones; and (iv) *engrailed* transcripts synthesized *in vitro* using *Drosophila* nuclear extracts initiate at the same sites (W.Soeller, T.Kornberg, unpublished).

If RNAs are transcribed from other portions of the locus, they escaped detection because of their scarcity or small size. Probes that could detect RNA present at less than 1 part per 10^5 failed to detect homologous sequences from the transcriptionally silent regions of the locus in preparations from 12 different developmental periods. Also, the undetected and low abundance RNAs would have to be transcribed from both sides of the demonstrated transcription unit and would necessarily have similar functions, inasmuch as lethal breakpoint *engrailed* mutations are members of a single complementation group, have similar phenotypes, and are located both 5′ and 3′ to the transcription unit. The finding that the structures of the three identified *engrailed* transcripts are unaffected by the outlying breakpoint mutations conclusively demonstrates that these RNAs are not derived from primary transcripts that originate in part from these regions. Rather, the influence of these breakpoint mutations on the pattern of *engrailed* expression (Weir and Kornberg, 1985) suggests that the genomic sequences outside of the transcription unit are involved in the temporal and spatial regulation of expression. Although it is possible that the breakpoint mutations alter *engrailed* function indirectly through the effects of the sequences they adjoin, the high density of transcription units in the *Drosophila* genome suggests instead that the transcriptionally silent, non-coding regions of the gene are indeed functional, perhaps containing multiple control elements that are responsible for the complex regulation of *engrailed* expression.

Drosophila has a relatively small genome, about 9×10^4 euchromatic kb. There are at least 5000 lethal complementation groups, suggesting that the average gene encompasses less than 15–20 kb (Judd and Young, 1973). The *engrailed* and *invected* genes, 70 and 35 kb, respectively, are considerably larger. Indeed, their unusual structure is evidence from comparisons of transcript density between the *engrailed* complex and the surrounding region; to either side of the complex, transcript density increases by an order of magnitude.

An earlier study of the 315-kb *Ace–rosy* segment of the *Drosophila* genome that contains an estimated 12 complementation groups also found distinct regions whose transcript densities differed by an order of magnitude (Hall *et al.*, 1983; Bossy *et al.*, 1984). Molecular, cytogenetic, and genetic analyses revealed approximately three transcripts per recessive lethal complementation group, a figure similar to the average of three poly(A)$^+$ RNAs per transcript group in the *engrailed* region. Although these calculations are imprecise, it is clear that genetic units in *Drosophila* can be closely juxtaposed, and that loci larger than 10–15 kb are not common.

Large *Drosophila* genes have been described include *Ultrabithorax* (Bender *et al.*, 1983), *Antennapedia* (Garber *et al.*, 1983; Scott *et al.*, 1983), *Notch* (Kidd *et al.*, 1983; Artavanis-Tsakonas, 1983), *achaete-scute* (Campuzano, 1985), as well as the two in the *engrailed* complex. The *Ubx*, *Antp*, *Notch*, and *invected* loci have small exons connected by very large introns, the intervening sequences contributing the major portion of the genes. In contrast, *achaete-scute* and *engrailed* have large regions upstream and downstream of the transcription units whose integrity is essential to their gene's function. For *engrailed*, expression during the pre-cellular syncytial blastoderm stages, when the interphase period is less than 5 min, mandates that the primary transcript be small (Karr *et al.*, 1985). Why so much DNA is needed outside of the transcription unit remains a mystery.

Materials and methods

Fly culture

Wild-type Oregon R flies were maintained in population cages at 25°C on standard cornmeal agar media supplemented with Baker's yeast. Prior to a timed collection, flies were provided with a pre-warmed, fresh and heavily-yeasted food plate to minimize retention of fertilized embryos. Embryos were collected on agar plates and aged at 25°C before harvesting. Larvae and pupae were from 6 h egg collections that were aged and collected in large plastic boxes containing 4 paper towels saturated with a yeast solution.

RNA preparation

Poly(A)$^+$ RNA was isolated from staged embryos, larvae, pupae and adults. Embryos were collected from agar plates in 0.4% Triton/7.0% NaCl, de-chorionated in 50% Chlorox (1.5 min), rinsed, and frozen in liquid nitrogen. Larvae, pupae and adults were frozen directly in liquid nitrogen. Frozen samples were ground to a powder with a mortar and pestle kept cool by dry ice and containing liquid nitrogen. Extracts were stored at −70°C. RNA was purified from the frozen powder essentially by the method of Chirgwin *et al.* (1979). The frozen extract was added to guanidinium buffer [5 M guanidinium isothiocyanate, 10 mM EDTA, 50 mM Hepes (pH 7.6), 5% 2-mercaptoethanol], and homogenized for 1 min in a Brinkman Polytron tissue solubilizer at top speed. Preparations of RNA not requiring poly(A) selection were not frozen and ground in liquid nitrogen but were homogenized in the guanidinium solution directly. The homogenate was clarified by centrifugation at 10 000 r.p.m. for 20 min at 5°C in a Beckman JA-17 rotor. The supernatant was brought to 4.0% sarkosyl with the addition of solid N-lauryl sarkosine (sodium salt), heated to 68°C for 10 min, layered over a 5.7 M CsCl cushion and pelleted for 20 h at 24 000 r.p.m. at 15°C in a Beckman SW27 rotor. RNA pellets were resuspended in 2 ml 8 M urea, 50 mM Hepes (pH 7.6) and 10 mM EDTA at 68°C for 5 min. An equal volume of 50 mM Hepes, 1 mM EDTA was added and the solution was extracted twice with phenol/chloroform/iso-amyl alcohol (24:24:1), twice with chloroform/iso-amyl alcohol (24:1) and twice with ether. Sodium acetate (2.5 M) was added to 0.25 M and the RNA was precipitated with two vol of 100% ethanol. Poly(A)$^+$ RNA was selected with oligo(dT) cellulose (Collaborative Research Inc., Type 2) according to Chirgwin *et al.* (1979). Total RNA was diluted to 2.5 mg/ml in diethylpyrocarbonate treated H_2O, heated to 68°C for 2 min and rapidly cooled on ice. The RNA solution (10 ml) was added to 5 ml cellulose and the 15 ml total brought to 0.5 M NaCl, 10 mM Hepes, 1 mM EDTA with a 10 × solution and agitated on a gyrorotary mechanical shaker for 2 h at room temperature. The cellulose was then washed 5 × by settling, removing the supernatant, and adding fresh binding buffer. The cellulose was then added to a sterile (10 ml) syringe, washed with 5 vol of 0.1 M NaCl, 10 mM Hepes, 1 mM EDTA wash buffer, and finally eluted with a 10 mM Hepes and 1 mM EDTA buffer without salt. This procedure yielded poly(A)$^+$ RNA preparations virtually free of ribosomal RNA.

Preparation of RNA from mutant embryos

Total RNA was prepared essentially according to the procedure of Cheley and Anderson (1984). De-chorionated embryos were homogenized using a 2 ml Teflon Dounce homogenizer in a total volume of 1 ml 7.8 M guanidine–HCl, 0.1 M potassium acetate. After 5–10 strokes, 0.6 vol of EtOH were added and the nucleic acid was allowed to precipitate overnight at −20°C. The precipitates were collected by a 10 min centrifugation at 12 000 *g* and resuspended in 0.4 ml DEPC treated 10 mM Hepes, pH 7.0, 1.0 mM EDTA. The nucleic acid was again precipitated with 2 vol EtOH for 1 h at −70°C with 60 μg *E. coli* tRNA/ml added as carrier. The pellets were resuspended in sample buffer (50% formamide, 2.2 M formaldehyde, 1 × Northern gel running buffer) and the RNA fractionated in agarose gels containing formaldehyde (see below).

Northern blots

RNA was fractionated on 2.2 M formaldehyde/0.9% agarose gels (Seakem; HGT, FMC) in 50 mM Hepes, 1 mM EDTA, 5 mM Na acetate, pH 7.0 (gel buffer). Samples were precipitated with ethanol, resuspended at 68°C for 10 min in gel buffer supplemented with 50% formamide, 2.2 M formaldehyde, bromophenol blue, and xylene cyanol. Each lane contained (as determined by A_{260}), 5 μg of poly(A)$^+$ or 20 μg of total RNA unless otherwise specified. RNA was transferred directly to either Millipore nitrocellulose (0.45 μm) or MSI nylon (0.45 μm) by the procedure of Thomas (1980) in 20 × SSPE for 12–18 h. The filters were rinsed in 2 × SSPE, dried, and baked for 2 h at 80°C *in vacuo*.

Northern blots were pre-hybridized in 50% formamide, 5 × SSPE, 10 × Denhardt's solution (0.2% Ficoll, 0.2% bovine serum albumin, 0.2% polyvinylpyrrolidone), 50 mM $NaPO_4$ (pH 6.5), and 250 μg/ml denatured and sonicated herring sperm DNA for 8 h at 42°C. Hybridization with probe was for 18–24 h at 42°C (or 36–60 h for single-stranded probes). The filters were washed twice in 2 × SSPE, 0.1% SDS and for 10 min at room temperature and 4 × in 0.1 × SSPE, 0.1% SDS, for 30 min at 50°C. The filters were dried and autoradiographed at −70°C using Kodak XAR-5 film with intensifying screens. Filters were reused 2–3 times according to Thomas (1980).

Hybridization probes

For all poly(A)$^+$ RNA blots, agarose gel-purified restriction fragments were labeled using a standard nick-translation reaction (Rigby *et al.*, 1977) in a 10 μl reaction containing 50 μCi of ^{32}P-labeled dCTP (New England Nuclear, 800 Ci/mmol). The reaction was stopped with the addition of 20 μl 200 mM EDTA and 50 μl of 10 mg/ml boiled, sonicated herring sperm DNA. The probe was purified by two passages through 750 μl P-10 (Bio-Rad) spin dialysis columns. The probe was boiled for 10 min before use; its specific activity was ~1.0–2.0 × 10^8 c.p.m./μg. For tissue RNA blots, more sensitive probes were required. Restriction fragments were cloned into M13 and a sequencing primer was used to prime the synthesis of a single stranded DNA (Hu and Messing, 1982). Ten nanograms of sequencing primer and 0.5 μg of single-stranded M13 template DNA were heated to ~90°C in a 10 μl volume containing 100 mM NaCl, 10 mM $MgCl_2$ and 20 mM Tris-HCl (pH 8.0), and allowed to cool. The cooled mixture was brought up to 20 μl volume containing 50 mM NaCl, 5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.0), 1 mM DTT, 50 μM unlabeled dCTP, dGTP, and TTP, 50 μCi of [^{32}P]dATP, 5 U of *E. coli* DNA polymerase Klenow fragment (Bethesda Research Labs) and incubated at 22°C for 1 h. This mixture was digested with a restriction endonuclease which cut within the polylinker region of the plasmid vector but not within the insert, spin dialyzed through P-10 and fractionated on a low melting temperature agarose (Sea Plaque, FMC) gel. The slice of agarose containing the labeled insert was identified by autoradiography, excised, added directly to hybridization solution, melted at 68°C for 10 min, filtered through Millex, and applied to the blot.

Acknowledgements

This work was supported by predoctoral training support from the NIH and the Weingart foundation to B.D. and by NIH and Weingart foundation grants to K.C., W.S., S.P. and T.K. We would like to thank for their kind gifts: Graeme Mardon for the RSV RNA, Phil Beachy for the *Ubx* cDNA clone, Matt Scott for the *Antp* and *fiz* clones, Michael Simon for the c-*src* cDNA clone, Sally Tobin and James Fristrom for the 5C actin clone, and D.Hudson Frew and Greg Stafford for help with the heraldic research which yielded the invected name. We also thank Larry Kauvar and Gail Martin for their comments on the manuscript.

References

Akam,M.E. (1983) *EMBO J.*, **2**, 2075–2084.
Artavanis-Tsakonas,S., Muskavitch,M.A.T. and Yedvobnick,B. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1977–1981.
Bender,W., Akam,M., Karch,F., Beachy,P.A., Peifer,M., Spierer,P., Lewis,E.B. and Hogness,D.S. (1983) *Science*, **221**, 23–29.
Bossy,B., Hall,L.M.C. and Spierer,P. (1984) *EMBO J.*, **3**, 2537–2541.
Campuzano,S., Carramolino,L., Cabrera,C.V., Ruiz-Gomez,M., Villares,R., Boronat,A. and Madolell,J. (1985) *Cell*, **40**, 327–338.
Cheley,S. and Anderson,R. (1984) *Anal. Biochem.*, **137**, 15–19.
Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry*, **18**, 5294–5299.
Coleman,K.G., Poole,S.J., Weir,M.P., Soeller,W. and Kornberg,T. (1987) *Genes Dev.*, **1**, 19–28.
Cory,S., Graham,M., Webb,E., Corcoran,L. and Adams,J.M. (1985) *EMBO J.*, **4**, 675–681.
Dente,L., Gianni,C. and Cortese,R. (1983) *Nucleic Acids Res.*, **11**, 1645–1655.
DiNardo,S., Kuner,J., Theirs,J. and O'Farrell,P. (1985) *Cell*, **43**, 59–66.
Fyrberg,E.A., Kindle,K.L., Davidson,N. and Sodja,A. (1980) *Cell*, **19**, 365–378.
Garber,R.L., Kuroiwa,A. and Gehring,W.J. (1983) *EMBO J.*, **2**, 2027–2036.
Garcia-Bellido,A. (1975) In *Cell Patterning*. CIBA Foundation Symposium, Amsterdam: North Holland, **29**, pp. 161–182.
Hall,L.M.C., Mason,P.J. and Spierer,P. (1983) *J. Mol. Biol.*, **169**, 83–96.
Hu,N. and Messing,J. (1982) *Gene*, **17**, 271–277.
Judd,B.H. and Young,M.W. (1973) *Cold Spring Harbor Symp. Quant. Biol.*, **38**, 573–579.
Karr,T.L., Drees,B., Ali,Z. and Kornberg,T. (1985) *Cell*, **43**, 591–601.
Kidd,S., Lockett,T.S. and Young,M.W. (1983) *Cell*, **34**, 421–433.
Kornberg,T. (1981a) *Proc. Natl. Acad. Sci. USA*, **78**, 1095–1099.
Kornberg,T. (1981b) *Dev. Biol.*, **86**, 363–372.
Kornberg,T., Siden,I., O'Farrell,P. and Simon,M. (1985) *Cell*, **40**, 45–53.
Kuner,J., Nakanishi,M., Ali,Z., Drees,B., Gustavson,E., Theis,J., Kauvar,L.M., O'Farrell,P. and Kornberg,T. (1985) *Cell*, **42**, 309–316.
Lawrence,P.A. and Morata,G. (1976) *Devl. Biol.*, **50**, 321–337.
Messing,J. and Viera,J. (1982) *Gene*, **19**, 269–283.
Poole,S.J., Kauvar,L.M., Drees,B. and Kornberg,T. (1985) *Cell*, **40**, 37–43.
Rigby,P., Dieckmann,M., Rhodes,C. and Berg,P. (1977) *J. Mol. Biol.*, **113**, 237–245.
Scott,M.P., Weiner,A.J., Hazelrigg,T.I., Polisky,B.A., Pirrota,V., Scalenghe,F. and Kaufman,T.C. (1983) *Cell*, **35**, 763–776.
Thomas,P.S. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5201–5205.
Tobin,S.L., Zulauf,E., Sanchez,F., Craig,E.A. and McCarthy,B.J. (1980) *Cell*, **19**, 121–131.
Weir,M. and Kornberg,T. (1985) *Nature*, **318**, 433–439.

Received on April 21, 1987; revised on June 19, 1987