The *invected* gene of *Drosophila*: sequence analysis and expression studies reveal a close kinship to the *engrailed* gene

Kevin G. Coleman, Stephen J. Poole, Michael P. Weir, Walter C. Soeller, and Thomas Kornberg

Department of Biochemistry, University of California at San Francisco, San Francisco, California 94143 USA

The *invected* and *engrailed* genes are juxtaposed in the *Drosophila* genome and are closely related in sequence and pattern of expression. The structure of the most abundant *invected* transcript was defined by obtaining the full-length cDNA sequence and by S1 nuclease sensitivity and primer extension studies; a partial sequence of the *invected* gene was determined; and the developmental profile of *invected* expression was characterized by Northern analysis and by in situ localization. The *invected* gene, like the *engrailed* gene, is expressed in the embryonic and larval cells of the posterior developmental compartments and in the embryonic hindgut, clypeolabrum, and nervous system. Like the *engrailed* gene, the *invected* gene can encode a protein of approximately 60 kD that contains a homeo box near its carboxyl terminus; indeed, a sequence of 117 amino acids in the carboxy-terminal region of both proteins is almost identical. The developmental role of the *invected* gene is not known.

[*Key Words*: *Drosophila*; *invected* gene; gene expression; sequence analysis; homeo box]

Received November 3, 1986; accepted November 25, 1986.

There are few examples of eukaryotic genes that are both functionally related and are organized as a linked unit, and it has been suggested that the apparently random distribution of genes on the chromosomes may be considered as an unspoken "third law" of Mendel (Lewis 1967). Several of the exceptional linked gene clusters have been described in *Drosophila*, principally the Bithorax Complex (Lewis 1978) and the Antennapedia Complex (Kaufman et al. 1980). Both clusters are composed of genes that have a role in regulating the segmental organization during embryogenesis and later development. Some of the constituent genes share sequences that suggest evolutionary kinship: A region of 180 nucleotides (homeo box) in genes of both complexes retains striking homology indicative of common ancestry, structure, and function (McGinnis et al. 1984).

The *engrailed* gene regulates the processes that subdivide the developing *Drosophila* embryo into reiterated segments and compartments (Kornberg 1981a). Therefore, it is similar to the genes of the Antennapedia and Bithorax Complexes in having a regulatory function during development. The *engrailed* gene has been cloned and the structural region that encodes a 60-kD nuclear protein has been defined (DiNardo et al. 1985; Kuner et al. 1985; Poole et al. 1985; T. Karr, N. Gay, and T. Kornberg, unpubl.). The *engrailed* gene also contains a homeo box sequence, similar to, yet distinctive from, those of the Antennapedia and Bithorax Complex genes (Poole et al. 1985). Although the two gene complexes are genetically linked, they are not linked to the *engrailed* gene.

Our previous studies on the structure of the genomic region containing the *engrailed* gene indicated the presence of a gene located just proximal to *engrailed* that also contains an *engrailed*-like homeo box (Poole et al. 1985). Previously termed *engrailed-related* and now designated the *invected* gene, this neighbor of the *engrailed* gene is shown here to contain a region of extensive homology with the *engrailed* gene that encompasses the homeo box and spans 117 amino acids. Furthermore, expression of both the *invected* and *engrailed* genes is limited to the same subset of embryonic and larval cells. We suggest that the *engrailed* and *invected* genes form a small, two-member complex involved in the organization and subdivision of the developing *Drosophila*.

Results

Isolation of invected *cDNA clones*

To identify sequences homologous to *engrailed*, portions of the *engrailed* cDNA c-2.4 (Poole et al. 1985) were hybridized with the DNA of recombinant phages that together contain 225 kb from the 48AB region of the *Drosophila* second chromosome (Kuner et al. 1985). Four

phages were detected with the cDNA probe: phages E7 and E8 (containing the *engrailed* transcription unit; Fig. 1A), phage E6 (containing the previously identified *engrailed-related* sequences), and phage E4 (previously not identified). Probes prepared from both phages E6 and E4 detected similarly sized transcripts in preparations of poly(A)$^+$ embryonic RNA. With a probe from phage E4, 21 cDNA clones were isolated which varied in size from 1.6 to 2.3 kb. Digestion of these clones with several restriction endonucleases suggested that they all shared common regions. Two clones (c-2.3 and c-2.1) were analyzed in detail.

Structure of the invected *cDNA clones c-2.3 and c-2.1*

When used to probe phages E1–E9, the *invected* cDNA clone c-2.3 hybridized to both phages E4 and E6 and to the *engrailed*-region phages E7 and E8. Both strands of cDNA c-2.3, portions of cDNA c-2.1, as well as much of the genomic DNA from the two *invected* regions, were sequenced.

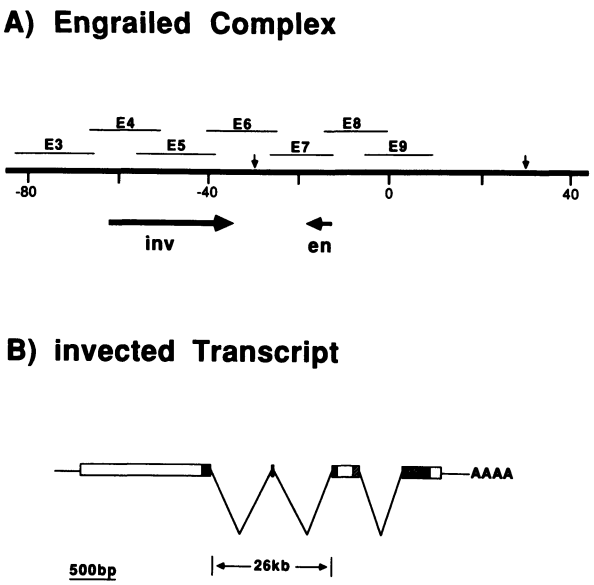

Figure 1. Structure and location of the *invected* gene. (*A*) Diagram of the *engrailed* complex. The centromere is to the left, and calibration is in kilobases, zero designating the insertion site of the *en*^{1} transposition (Kuner et al. 1985). Vertical arrows indicate the most proximal and the most distal breakpoints of lethal *engrailed* mutations and the horizontal arrows the positions of the transcribed regions in the *invected* and *engrailed* genes. Genomic regions within individual phage (E3–E9) are shown above their relevant locations. (*B*) Structure of the *invected* cDNA clone c-2.3. Direction of transcription is left to right, as indicated by the *inv* arrow in *A*. Thin horizontal lines indicate untranslated regions; boxes indicate the 1740-nucleotide open reading frame. Lengths are to scale, except for the 26 kb spanned by the first and second introns; the exons have 1588, 6, 161, and 694 nucleotides and the third intron has 417 nucleotides. Sequences represented by open boxes are *invected* specific, by hatched box are the homeo box, and by filled boxes are the non-homeo box but *engrailed*-related portions.

The cDNA portion of clone c-2.3 contains 2288 bp and a poly(dA) : (dT) tail at one end. Northern analysis with strand-specific probes made from the cDNA revealed that the poly(dA) tail was at the 3′ end of *invected* transcripts (data not shown). Clone c-2.1 extensively overlaps c-2.3, and differs in lacking approximately 200 bp and a poly(dA) tail and in extending further in the opposite direction by 26 bp. As demonstrated below, together these clones represent the full length of the most abundant *invected* transcript.

Comparison of the cDNA sequences with two corresponding genomic regions from phages E4 and E6 revealed 17 nucleotide differences. Thirteen of these differences would not change the putative coding sequence of the cDNA's open reading frame (see below) and may be attributed to polymorphisms (the genomic and cDNA libraries were constructed from flies with different backgrounds); four changed the presumptive coding capacity (Gly-39 → Val, Cys-45 → Val, Val-81 → Ala, and Met-159 → Leu). Three points of non-colinearity were revealed, indicative of three intervening sequences in the primary transcript. The splice junctions for these intervening sequences are located at positions 1588, 1594, and 1755 (position 1 is the most proximal nucleotide of c-2.1). The first exon (1588 bp in length) is derived from the *invected* genomic region that is contained within phage E4 and is 43 kb proximal to and downstream from the 3′ end of the *engrailed* transcription unit (Fig. 1B). The first and second intervening sequences together span approximately 26 kb, within which is located the 6-bp Exon II. As the precise location of Exon II has not been determined, the relative sizes of the two introns bordering it are unknown. Exon II was present in all six cDNAs that were sequenced; the possibility that it represents a polymorphism in the fly population from which the cDNA library was generated cannot be excluded. The 3′ end of the *invected* transcript (Exons III and IV, which are 161 and 694 bp in length, respectively) is derived from the genomic region that is contained within phage E6 and is located 17 kb proximal to the 3′ end of the *engrailed* transcription unit. A 417-nucleotide intervening sequence interrupts Exons III and IV, as previously described (Poole et al. 1985).

The most extensive open reading frame in the cDNAs has 1740 bp (nucleotides 283–2022) and is in frame with the homeo box sequence. Translation of the 1740-bp open reading frame from the position of its first ATG codon (295) would produce a protein with 576 amino acids and a molecular mass of 60,964 daltons (Fig. 2). There are eight additional in-frame ATG codons located 5′ of the homeo box sequence. None of the nine ATG codons has a significant degree of homology to a general eukaryotic initiation consensus sequence (Kozak 1984), although the second (amino acid 41 in Fig. 2) and eighth (amino acid 212) ATG codons have the best homology with other *Drosophila* initiation sites [NNNANN(C/A)A(A/C)(A/C)ATG; D. Cavener, pers. comm.]. There are eight other upstream ATG codons that are out of frame, but none would produce a polypeptide with more than 24 amino acids.

```
TCGACGTCGGCAGAGCGAAACGATAAGAGTTCGCGAGAGAATCAGTTGGTCGGCGGTTCGTAACAGCGCACGTTACGTTTTTAACGGACGAGCGGATTAACCTCGGTATCGACAAGTTT  119

ATCACCCAACCCAAAAACCAATTGTCAATCGAAGACGATAGCGCGGCGTATTTGCCAATCGAAGTGATATCTTTTGGCTGTGGTTAATGTTTTTAATGCTTTCTGGCTGCCCTGAAATG  238


                                                  1                                   10
                                                  Met Ser Thr Leu Ala Ser Thr Arg Pro Pro Pro Leu Lys Leu Thr
CCGTCGGCTTTTGGTCCCTGTGAACTTTCGAGCTGAAGACCTAAGCCCTCAAACTG  ATG TCC ACC TTG GCC AGC ACT CGA CCG CCC CCG CTC AAG CTG ACT  339

                20                                      30                                      40
Ile Pro Ser Leu Glu Glu Ala Glu Asp His Ala Gln Glu Arg Arg Ala Gly Gly Gly Gly Gln Glu Val Gly Lys Met His Pro Asp Cys
ATT CCT TCT TTG GAG GAG GCA GAG GAC CAC GCG CAG GAG AGG AGA GCG GGT GGA GGT GGC CAG GAA GTG GGT AAA ATG CAT CCG GAT TGC  429

                50                                      60                                      70
Leu Pro Leu Pro Leu Val Gln Pro Gly Asn Ser Pro Gln Val Arg Glu Glu Glu Glu Asp Glu Gln Thr Glu Cys Glu Glu Gln Leu Asn
TTG CCC TTG CCG CTT GTC CAG CCG GGC AAC TCG CCC CAA GTG CGA GAG GAG GAG GAG GAC GAG CAG ACG GAG TGC GAG GAG CAG CTG AAC  519

                80                                      90                                      100
Ile Glu Asp Glu Glu Val Glu Glu Glu His Asp Leu Asp Leu Glu Asp Pro Ala Ser Cys Cys Ser Glu Asn Ser Val Leu Ser Val Gly
ATA GAG GAT GAA GAG GTC GAG GAG GAG CAC GAC TTG GAC CTG GAG GAT CCA GCC AGC TGC TGC AGT GAG AAT AGT GTC TTG AGT GTG GGC  609

                110                                     120                                     130
Gln Glu Gln Ser Glu Ala Ala Gln Ala Ala Leu Ser Ala Gln Ala Gln Ala Arg Gln Arg Leu Leu Ile Ser Gln Ile Tyr Arg Pro Ser
CAG GAA CAA TCG GAA GCA GCA CAG GCT GCG CTA TCC GCT CAA GCT CAG GCC AGG CAG AGG CTT TTG ATC AGC CAA ATC TAT CGA CCG TCG  699

                140                                     150                                     160
Ala Phe Ser Ser Thr Ala Thr Thr Val Leu Pro Pro Ser Glu Gly Pro Pro Phe Ser Pro Glu Asp Leu Met Gln Leu Pro Pro Ser Thr
GCA TTC AGC AGC ACT GCT ACC ACT GTT TTG CCG CCC AGC GAA GGT CCG CCC TTC TCA CCC GAA GAT CTC ATG CAA CTG CCT CCC TCC ACT  789

                170                                     180                                     190
Gly Thr Phe Gln Glu Glu Phe Leu Arg Lys Ser Gln Leu Tyr Ala Glu Glu Leu Met Lys Gln Gln Met His Leu Met Ala Ala Ala Arg
GGA ACC TTC CAG GAG GAG TTC CTG CGA AAG TCC CAG CTG TAC GCC GAG GAG CTG ATG AAA CAG CAG ATG CAT CTA ATG GCC GCC GCC AGG  879

                200                                     210                                     220
Val Asn Ala Leu Thr Ala Ala Ala Ala Gly Lys Gln Leu Gln Met Ala Met Ala Ala Ala Ala Val Ala Thr Val Pro Ser Gly Gln Asp
GTG AAC GCT CTC ACG GCA GCG GCG GCG GGA AAA CAG CTG CAA ATG GCA ATG GCG GCG GCG GCG GTT GCC ACA GTG CCC AGC GGT CAG GAT  969

                230                                     240                                     250
Ala Leu Ala Gln Leu Thr Ala Thr Ala Leu Gly Leu Gly Pro Gly Gly Ala Val His Pro His Gln Gln Leu Leu Leu Gln Arg Asp Gln
GCA CTT GCT CAG CTG ACG GCC ACG GCA TTG GGT CTG GGA CCC GGT GGA GCG GTG CAT CCC CAT CAG CAA TTG TTA CTG CAA AGG GAT CAG 1059

                260                                     270                                     280
Val His His His His His Met Gln Asn His Leu Asn Asn Asn Glu Asn Leu His Glu Arg Ala Leu Lys Phe Ser Ile Asp Asn Ile Leu
GTC CAC CAC CAT CAT CAC ATG CAG AAT CAC CTG AAT AAC AAT GAA AAT CTG CAC GAA AGG GCG CTC AAA TTC AGC ATA GAC AAT ATC CTG 1149

                290                                     300                                     310
Lys Ala Asp Phe Gly Ser Arg Leu Pro Lys Ile Gly Ala Leu Ser Gly Asn Ile Gly Gly Gly Ser Val Ser Gly Ser Ser Thr Gly Ser
AAG GCG GAC TTT GGT TCC AGA TTG CCC AAG ATT GGT GCT TTA AGT GGC AAT ATC GGT GGT GGC AGT GTA AGT GGG AGC AGT ACT GGC AGC 1239

                320                                     330                                     340
Ser Lys Asn Ser Gly Thr Thr Asn Gly Asn Arg Ser Pro Leu Lys Ala Pro Lys Lys Ser Gly Lys Pro Leu Asn Leu Ala Gln Ser Asn
AGT AAA AAT TCT GGA ACT ACC AAT GGC AAC AGA TCT CCC CTA AAG GCG CCC AAG AAG TCG GGA AAG CCA CTG AAT CTG GCT CAA AGT AAC 1329

                350                                     360                                     370
Ala Ala Ala Asn Ser Ser Leu Ser Phe Ser Ser Ser Leu Ala Asn Ile Cys Ser Asn Ser Asn Asp Ser Asn Ser Thr Ala Thr Ser Ser
GCC GCC GCC AAC TCG AGC TTG AGC TTC TCC AGC TCG CTG GCG AAT ATA TGC AGC AAC AGC AAC GAT TCC AAT AGC ACC GCC ACC AGC AGC 1419

                380                                     390                                     400
Ser Thr Thr Asn Thr Ser Gly Ala Pro Val Asp Leu Val Lys Ser Pro Pro Pro Ala Ala Gly Ala Gly Ala Thr Gly Ala Ser Gly Lys
AGC ACC ACC AAC ACC TCG GGG GCA CCA GTG GAT CTG GTC AAG TCG CCA CCG CCA GCG GCA GGA GCG GGT GCC ACA GGA GCA TCC GGT AAA 1509

                410                                     420                                     430
Ser Gly Glu Asp Ser Gly Thr Pro Ile Val Trp Pro Ala Trp Val Tyr Cys Thr Arg Tyr Ser Asp Arg Pro Ser Ser Gly Arg Ser Pro
TCG GGC GAG GAT TCC GGC ACT CCC ATC GTT TGG CCA GCG TGG GTC TAC TGC ACT CGC TAC AGC GAT CGT CCC AGC TCA GGT CGA AGT CCA 1599
                                                                                                         ▲       ▲
                440                                     450                                     460
Arg Ala Arg Lys Pro Lys Lys Pro Ala Thr Ser Ser Ser Ala Ala Gly Gly Gly Gly Gly Gly Val Glu Lys Gly Glu Ala Ala Asp Gly
CGA GCG CGA AAG CCC AAG AAG CCG GCG ACG TCC AGT TCG GCG GCA GGT GGT GGG GGT GGG GGC GTC GAG AAG GGG GAG GCC GCC GAT GGG 1689

                470                                     480                                     490
Gly Gly Val Pro Glu Asp Lys Arg Pro Arg Thr Ala Phe Ser Gly Thr Gln Leu Ala Arg Leu Lys His Glu Phe Asn Glu Asn Arg Tyr
GGC GGG GTG CCG GAG GAC AAA AGG CCG CGA ACG GCC TTC AGC GGA ACG CAG TTG GCC AGA CTG AAG CAC GAG TTC AAC GAG AAT CGC TAT 1779
                                                                                       ▲
                500                                     510                                     520
Leu Thr Glu Lys Arg Arg Gln Gln Leu Ser Gly Glu Leu Gly Leu Asn Glu Ala Gln Ile Lys Ile Trp Phe Gln Asn Lys Arg Ala Lys
CTG ACG GAG AAG CGA CGC CAG CAG CTG AGC GGG GAA CTG GGA CTG AAC GAG GCG CAG ATC AAG ATC TGG TTC CAG AAC AAA CGG GCC AAG 1869

                530                                     540                                     550
Leu Lys Lys Ser Ser Gly Thr Lys Asn Pro Leu Ala Leu Gln Leu Met Ala Gln Gly Leu Tyr Asn His Ser Thr Ile Pro Leu Thr Arg
CTG AAA AAG TCG AGC GGC ACC AAG AAT CCG CTG GCG CTG CAG CTG ATG GCG CAG GGA TTG TAC AAC CAC TCG ACG ATA CCG CTG ACC CGC 1959

                560                                     570                         576
Glu Glu Glu Glu Leu Gln Glu Leu Gln Glu Ala Ala Ser Ala Arg Ala Ala Lys Glu Pro Cys AM
GAG GAG GAG GAG CTG CAG GAG CTG CAG GAG GCG GCT AGT GCC CGT GCC GCC AAG GAG CCC TGC TAG AAGGAGGTGGCGTGTGCGGCAATATCTACAA 2056

TCTAGTATTTATGGAGTAGTGTGTAAGCTAGCTTTAGAAATTCTGAGCGATTAAGTTGTACAATATTCTAGTCCGCCGCACACGCTAGTCGAACGCAGAATCAGGTGAAGAAATCTTCC 2175

GGAGAATTGGCCGCTGGGCGTAGAAATCCCCCGAAGAGCTATGATTGTGTGTGCGTTTTGTATAATAGATTCATAAATTCACAACTAAATAATTTAATAAACATTTAAATTATAAAAAA 2294
```

Figure 2. Sequence and translation of the *invected* cDNA clone c-2.3. Both strands were sequenced and the translation shown is of the 1740-nucleotide open reading frame starting with the first ATG at nucleotide 295. Amino acids are numbered starting with this ATG. The homeo box is underlined. The splice junctions of the three introns are at nucleotides 1588, 1594, and 1755 (arrowheads). (Clone c-2.3 actually starts at nucleotide 27; the first 26 nucleotides shown are derived from clone c-2.1).

Dispersed among the first 470 amino acids of the protein sequence of the largest open reading frame are sequences consisting of three to five consecutive residues of either glycine (occurring four times), alanine (four times), glutamic acid (three times), serine (three times), proline (twice), or histidine, leucine, or asparagine (once each). Longer poly (amino acid) sequences (such as those found in the *engrailed* sequence; Poole et al. 1985) are not present. Near the carboxyl terminus (residues 470–529) is the homeo box sequence, similar to the homeo box sequences in the *Drosophila* Antennapedia (30 of 60 amino acids are identical) and *fushi tarazu* (28 of 60 amino acids are identical) genes, but most homologous to the *engrailed* homeo box (52 of 60 amino acids are identical).

Homologous sequences in the invected *and* engrailed *genes*

The potential protein-coding sequences of the *invected* and *engrailed* genes are similarly sized (1740 and 1656 nucleotides, respectively) and share striking homologies in their 3′ portions, but no apparent homology elsewhere. As noted previously (Poole et al. 1985), the homeo boxes of the two genes have almost the same sequence, and five of the eight differences represent conservative amino acid substitutions. Also, the genomic sequences for both homeo boxes are interrupted by intervening sequences at the same location, although sequence comparisons indicate that these intervening sequences are not homologous (Poole et al. 1985).

Sequence homologies between the *invected* and *engrailed* genes extend beyond the homeo boxes in both directions (Fig. 3). Seventy-eight (84%) of the 93 bases and 26 (84%) of the 31 amino acid residues (all five differences are conservative substitutions) immediately downstream from the homeo boxes are identical. Beyond this, the *invected* gene encodes 16 amino acids before a stop codon is reached, and the *engrailed* gene encodes seven. There is a second region of homology upstream of the homeo box. It begins with amino acid residues 416 of *invected* and 422 of *engrailed* and extends for 17 amino acids (all identical) before intervening sequences interrupt both genes. The *engrailed* intron has 1133 nucleotides (S. Poole and T. Kornberg, unpubl.). In contrast, the *invected* gene has two intervening sequences that together span approximately 26,000 DNA bp. The six-nucleotide Exon II (encoding two amino acids) between these two introns has no counterpart in the *engrailed* gene. A stretch of nine amino acid residues, of which six are identical, follows these splice junctions and concludes the upstream region of homology. Five amino acid residues of the *engrailed* gene and 26 of the *invected* gene separate this region from their respective homeo boxes.

The 5′ end of the invected *transcript*

To define the start site of the *invected* transcript, a 539-nucleotide single-stranded fragment (nucleotides −352

```
                    420                                 430                                 440
INV  trp pro ala trp val tyr cys thr arg tyr ser asp arg pro ser ser gly arg ser pro arg ala arg lys pro lys lys pro ala thr
      *   *   *   *   *   *   *   *   *   *   *   *   *   *   *           *   *       *       *   *   *       *
EN   trp pro ala trp val tyr cys thr arg tyr ser asp arg pro ser ser gly   .   .  pro arg tyr arg arg pro lys gln pro lys asp

                    450                                 460                                 470
INV  ser ser ser ala ala gly gly gly gly gly gly val glu lys gly glu ala ala asp gly gly gly val pro glu asp lys arg pro arg
                                                                                                    *   *   *   *
EN   lys thr asn  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   . asp glu lys arg pro arg

                    480                                 490                                 500
INV  thr ala phe ser gly thr gln leu ala arg leu lys his glu phe asn glu asn arg tyr leu thr glu lys arg arg gln gln leu ser
      *   *   *   *           *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
EN   thr ala phe ser ser glu gln leu ala arg leu lys arg glu phe asn glu asn arg tyr leu thr glu arg arg arg gln gln leu ser

                    510                                 520                                 530
INV  gly glu leu gly leu asn glu ala gln ile lys ile trp phe gln asn lys arg ala lys leu lys lys ser ser gly thr lys asn pro
          *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
EN   ser glu leu gly leu asn glu ala gln ile lys ile trp phe gln asn lys arg ala lys ile lys lys ser thr gly ser lys asn pro

                    540                                 550                                 560
INV  leu ala leu gln leu met ala gln gly leu tyr asn his ser thr ile pro leu thr arg glu glu glu glu leu gln glu leu gln glu
      *   *   *   *   *   *   *   *   *   *   *       *   *   *   *       *   *   *   *   *   *
EN   leu ala leu gln leu met ala gln gly leu tyr asn his thr thr val pro leu thr lys glu glu glu glu leu glu met arg met asn

                    570
INV  ala ala ser ala arg ala ala lys glu pro cys STOP

EN   gly gln ile pro STOP
```

Figure 3. The *invected* and *engrailed* coding regions have extensive homologies. Translations of the *invected* and *engrailed* cDNA sequences have been aligned in the regions of homology. Numbers indicate the relative amino acid positions in the two sequences, asterisks the positions of identity, and the line the location of the homeo box. Two gaps (dots) have been inserted into the *engrailed* sequence to maintain homology.

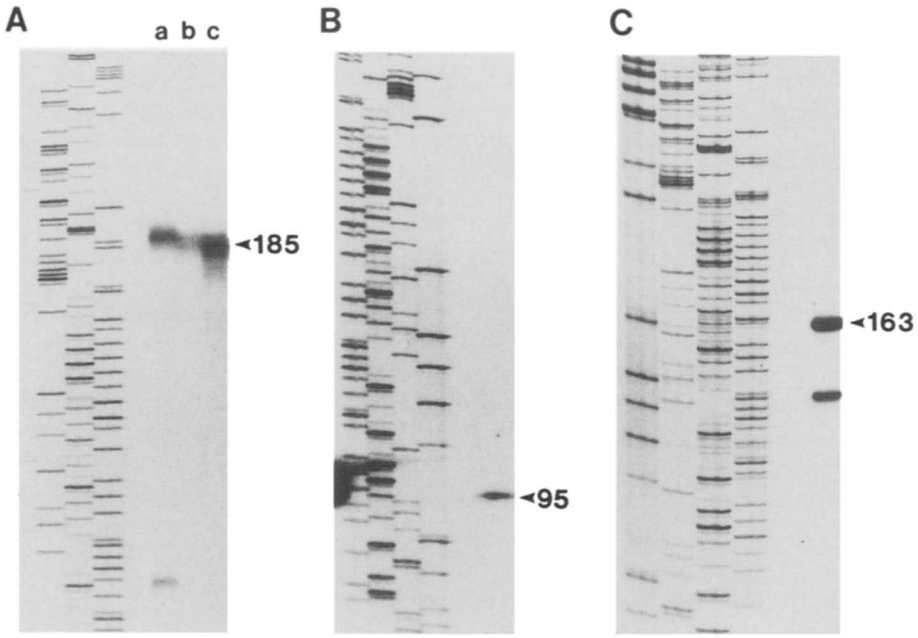

Figure 4. S1 nuclease and primer extension mapping of the 5′ end of the *invected* transcript. (*A*) A 539-nucleotide uniformly labeled and single-stranded genomic fragment (extending to nucleotide 187 of Fig. 2) was hybridized with 0- to 12-hr poly(A)$^+$ RNA. The RNA–DNA hybrid molecules were digested with either 26.5 (lane *a*), 53 (lane *b*), or 106 (lane *c*) units of S1 nuclease. Increasing amounts of enzyme reduce the size of the protected fragment to approximately 187 bases. An unrelated sequencing ladder was used for size calibration. (*B* and *C*) End-labeled synthetic DNA primers 1 and 2, complementary to nucleotides 72–95 and 138–163, respectively, were hybridized to 0- to 12-hr poly(A)$^+$ RNA and extended with reverse transcriptase. Hybrid fragments were fractionated electrophoretically and their sizes were calibrated with an unrelated sequencing ladder. (*B*) Primer 1; (*C*) primer 2.

to +187; Fig. 2), uniformly labeled and spanning the presumptive start site, was annealed to poly(A)$^+$ embryonic RNA. The hybrid molecules were digested by S1 nuclease and were fractionated electrophoretically (Fig. 4A). The protected fragment (~187 nucleotides) indicates either a transcription start or splice site around position 1.

To determine if the 5′ end of the *invected* gene was located at this site, two synthetic oligonucleotides (primer 1 and primer 2, homologous to nucleotides 72–95 and 138–163 of the cDNA, respectively) were separately end-labeled with [^{32}P]ATP and polynucleotide kinase and were annealed to poly(A)$^+$ embryonic RNA. These oligomers were extended with reverse transcriptase and the reaction products were fractionated electrophoretically. Both primers were extended to position 1 (Fig. 4B,C). Therefore, position 1 is likely to be the site of initiation for *invected* gene transcription.

The invected *transcripts*

Genomic DNA probes prepared from the centromere distal *invected* region hybridize to three transcripts, 3.4, 2.7, and 1.2 kb in length (Poole et al. 1985). To determine which of these transcripts corresponds to the mature *invected* mRNA, a Northern blot of 3–12 hr poly(A)$^+$ RNA was hybridized with the *invected* cDNA clone c-2.3 (Fig. 5, lane a). This probe hybridized strongly to the 2.7-kb transcript but only weakly to the 3.4- and 1.2-kb transcripts. A smaller probe that was *invected*-specific (nucleotides 1–1341), lacking sequences homologous to *engrailed*, detected the same pattern of transcripts (Fig. 5, lane b). However, when a similar Northern blot was hybridized with an *invected* intron-specific probe (a 438-nucleotide fragment from the 3′ region of the second intron), only the 3.4- and 1.2-kb transcripts were observed (Fig. 5, lane c). This intron-specific probe was used to isolate two cDNA clones. Sequence analysis revealed that both clones contain sequences from the 3′ end of the second intron, contain Exons III and IV, but lack the third intron. Since, in addition, the second intron sequences contained many stop codons in all possible reading frames, we conclude that the 2.7-kb transcript is the mature *invected* message, and the 3.4- and 1.2-kb transcripts represent other less abundant products of its maturation process.

In situ localization of invected *transcripts*

Genetic evidence has shown that the cells of the posterior compartment of each segment require *engrailed* function for normal development (Lawrence and Morata

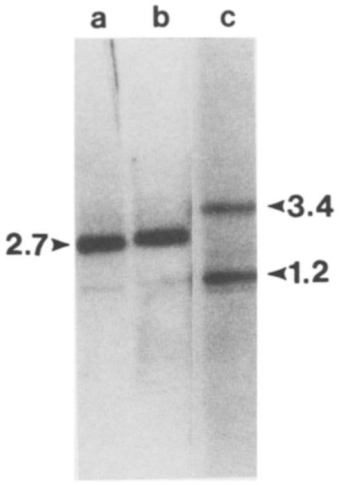

Figure 5. Expression of the *invected* gene. Poly(*A*)$^+$ RNA (10μg) from 0- to 12-hr embryos was fractionated on a formaldehyde-agarose gel, transferred to nitrocellulose, and hybridized with *invected* cDNA and genomic probes as described previously (Poole et al. 1985). Probes were: (lane *a*) entire insert of cDNA c-2.3; (lane *b*) a 1315-bp *Eco*RI–*Xho*I fragment of c-2.3 (nucleotides 27–1341) containing only sequences unique to *invected*; (lane *c*) a single-stranded probe complementary to a 438-base region near to the 3′ end of the second intron.

1976; Kornberg 1981a,b; Lawrence and Struhl 1982). In addition, in situ hybridization of an *engrailed* cDNA probe to frozen sections of embryos and larvae and to whole imaginal discs has demonstrated that the *engrailed* transcripts are expressed specifically in cells of the posterior compartments (Kornberg et al. 1985). Using in situ hybridization, *engrailed* transcripts were detected during cellularization of the blastoderm as annuli of expressing cells, initially in a pattern with a repeat length of four segments, then two segments, and then one segment (Weir and Kornberg 1985). Midway through germ-band elongation, the *engrailed* locus is expressed in 14–15 reiterated, evenly spaced bands. Expression was also observed in the clypeolabrum and hindgut after the completion of germ-band elongation, and in the neural ganglia by the time of germ-band shortening.

To compare directly the patterns of *invected* expression to those of *engrailed*, alternate serial sections from frozen embryos of various ages were hybridized with cDNA probes specific to *invected* or *engrailed*. We demonstrate here that whereas after germ-band elongation the patterns of *engrailed* and *invected* transcripts are indistinguishable, the levels of *invected* transcripts are significantly lower than those of *engrailed* before the completion of germ-band elongation. Virtually no hybridization of the *invected* probe was detected until midway through germ-band elongation, when a relatively low number of autoradiographic grains was observed (Fig. 6). Between mid- and complete germ-band elongation, grains formed reiterated patterns with apparent repeat lengths of two or four segments (not shown). After germ-band elongation, *invected* transcripts were found in segmental patterns similar to those of *engrailed* transcripts. Comparision of serial sections that had been hybridized with probes specific for either gene indicated that at germ-band elongation, cells in the same part of each segment expressed both genes simultaneously. The intensities of hybridization of the *engrailed* and *invected* probes were also similar (Fig. 7b,c). By germ-band shortening, *invected*, like *engrailed*, was expressed at the anterior edge of the deep segmental grooves (Fig. 8a), in domains considered to be posterior compartments (Kornberg et al. 1985). Strong expression of both *invected* and *engrailed* was also observed in the clypeolabrum, hindgut, and the neural ganglia (Figs. 7a–c and 8b). In summary, although the embryonic developmental period during which the patterns of *invected* expression develop is later than for *engrailed*, the sequence of patterns in which the two genes are expressed is similar, and both appear to be expressed in the posterior compartment cells.

The patterns of *invected* and *engrailed* expression were also compared in sections of frozen third instar larvae and in whole mounts of wing imaginal discs. Imaginal discs are nests of cells that grow in the larval body and later secrete the adult integument during metamorphosis. Of the different discs, the wing imaginal disc of

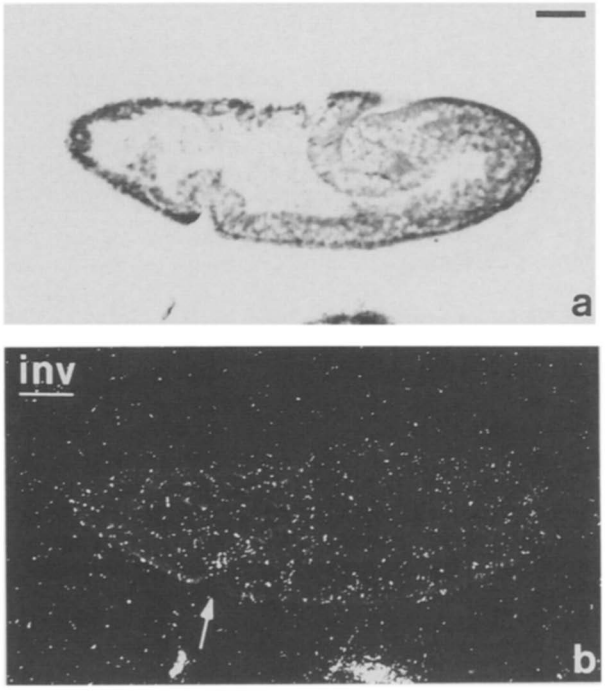

Figure 6. *invected* transcripts in gastrulating embryos. Bright-field (*a*) and corresponding dark-field (*b*) micrographs were recorded after hybridization with an *invected*-specific cDNA probe. The embryo is oriented anterior left, dorsal up. Sparse concentrations of grains are near the cephalic furrow (arrow). Embryos of this age have strong *engrailed* expression in every segment (Weir and Kornberg 1985). Scale bar, 50μm.

the third larval instar has been particularly well characterized, in that fate mapping (Bryant 1975) and clonal analysis (Brower et al. 1981) have precisely located its anterior and posterior compartments.

Larval sections and fixed imaginal discs were hybridized to either an *invected* or *engrailed* gene probe, and homologous RNA was quantitated by autoradiography. For both probes in both preparations, grains were detected over only half of the wing imaginal disc (Fig. 9a–c), in a region corresponding to the location of the posterior compartment (Brower et al. 1981; Kornberg et al. 1985). We conclude that the *invected* gene is expressed in the cells of the posterior, but not the anterior, compartment of the wing imaginal disc.

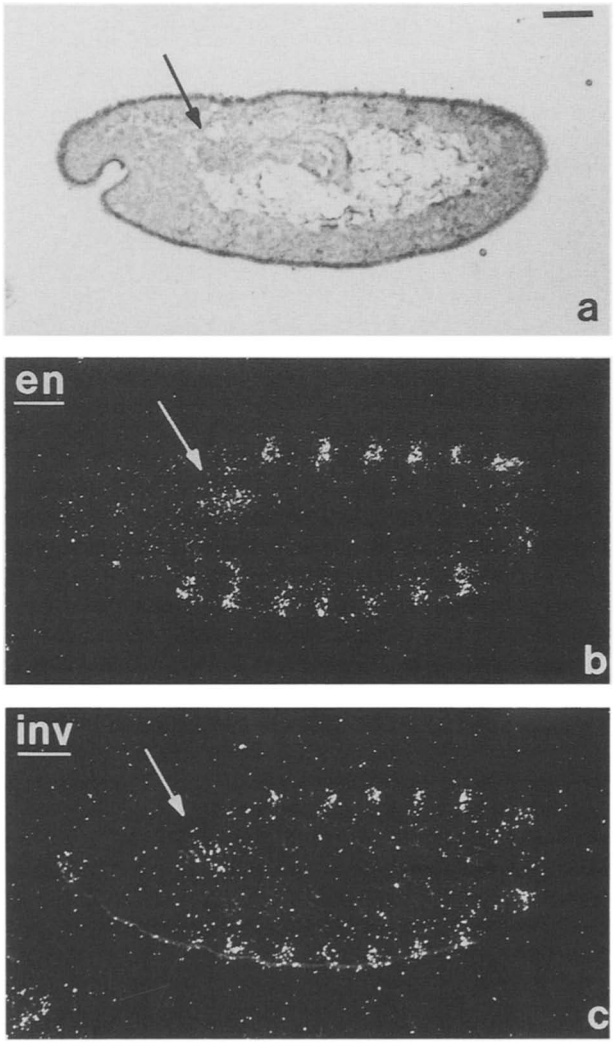

Figure 7. *invected* and *engrailed* transcripts in germ-band elongated embryos. Micrographs of serial sections were hybridized with cDNA probes specific for transcripts of *engrailed* (*a*, bright-field; *b*, dark-field) and invected (*c*, dark-field). Grain concentrations are similar for both probes over the ectoderm and hindgut (arrow). Orientation is anterior left, dorsal up. Scale bar, 50 μm.

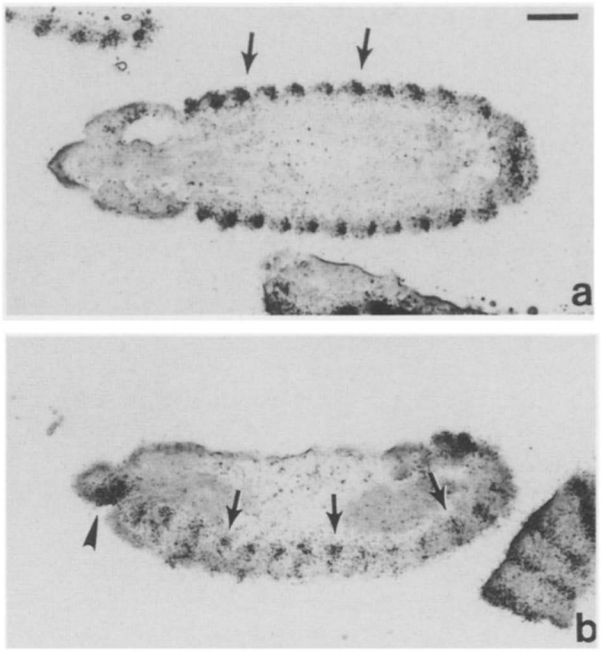

Figure 8. *invected* transcripts in germ-band shortened embryos. Bright-field micrographs of embryo sections after hybridization with an *invected* probe. Autoradiographic grains are localized in *a* over the posterior portion of each segment (arrows) in this frontal section oriented anterior left; and in *b* over part of the clypeolabrum (arrowhead) and the neural ganglia (three are indicated with arrows) in this sagittal section oriented anterior left, dorsal up. Scale bar, 50 μm.

Discussion

In the polytene region 48A of the second chromosome of *Drosophila* are two genes, *engrailed* and *invected*, that span a minimum of 70 kb (defined by *engrailed* breakpoint mutations) and 30 kb (defined by the *invected* transcription unit), respectively. They are immediately juxtaposed, without any interposed transcription units (B. Drees, Z. Ali, and T. Kornberg, in prep.). They share some features that are conserved to a striking degree, and yet they also differ. It is not known how they are related functionally nor is the significance of their close proximity understood.

Features that are similar include portions of the sequences of the proteins they can encode and the patterns of their expression during development. Translation of the *engrailed* transcript appears to initiate at the first in-frame ATG of its largest open reading frame (T. Karr, N. Gay, S. Poole, and T. Kornberg, unpubl.), and to produce a protein of 59,430 daltons. The *engrailed* protein has several long stretches of polyglutamine, polyglutamic/aspartic acid, and polyserine in its amino-terminal half and a homeo box in its carboxy-terminal portion (Poole et al. 1985). The largest open reading frame of the *invected* gene can encode a protein of molecular weight 60,964, and for this discussion, we assume that *Drosophila* produces this *invected* product. The protein

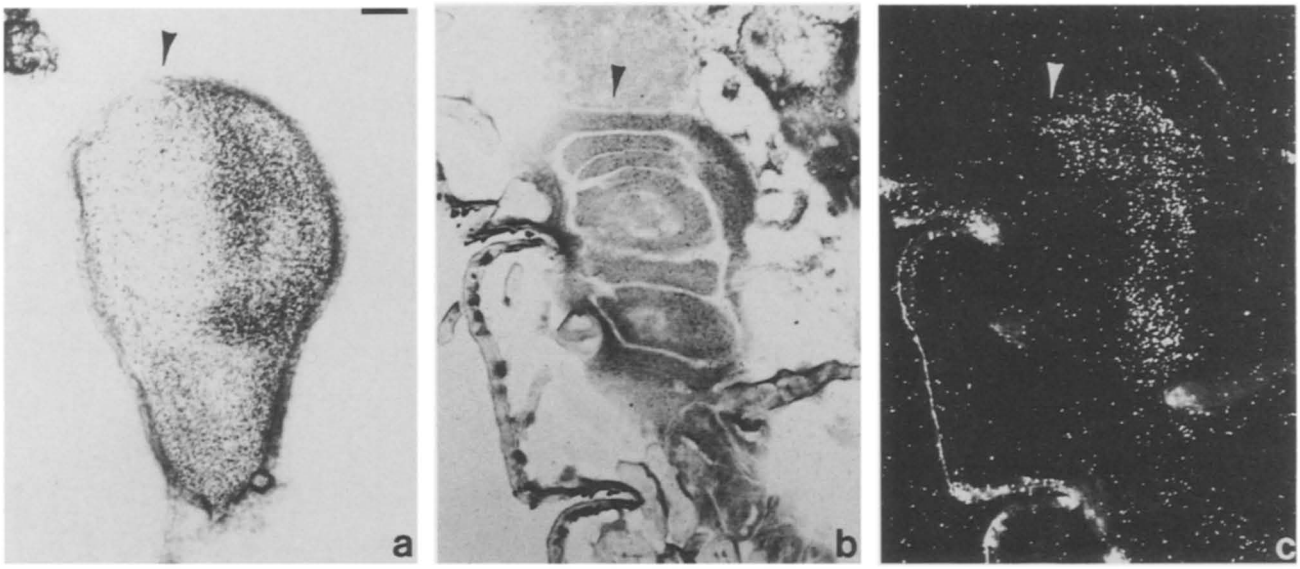

Figure 9. *invected* transcripts in imaginal wing discs. Micrographs of third instar wing imaginal discs, whole mounts (*a*) or sectioned (*b,c*). The discs are oriented anterior left, ventral up. Discs were hybridized with a cDNA probe specific for *invected*. In both discs, high concentrations of grains are over the posterior compartment. Note that the anterior–posterior compartment border extends anteriorly at the ventral edge of the disc (arrowheads), consistent with indications from clonal analysis (Brower et al. 1981). Scale bar, 50 μm.

sequence has numerous short poly(amino acid) stretches in its amino-terminal portion, and a homeo box near its carboxyl terminus. The *invected* homeo box is a virtual copy of the *engrailed* homeo box and is within a 117-amino-acid region that is almost identical to the homologous part of the *engrailed* protein. These strong similarities within the conserved regions of *engrailed* and *invected* suggest structural and therefore functional identity of approximately 13,700-dalton portions of the respective carboxyl termini. No apparent homologies exist elsewhere within the protein sequences.

Both *invected* and *engrailed* are expressed selectively in the same subset of embryonic and larval cells: in the cells of the posterior developmental compartments and in the developing embryonic ganglia, hindgut, and clypeolabrum. Neither gene is expressed in adult flies (B. Dress, Z. Ali, and T. Kornberg, unpubl.). Indeed, a comparison of the sequence of patterns that characterize their expression reveals a difference only in timing. Whereas stripes of *engrailed*-expressing cells were observed at the onset of cellularization, *invected* stripes were not evident until mid-germ-band elongation. After germ-band elongation, stripes of posterior compartment cells express both genes. However, this difference in timing may be more apparent than real. Although the later detection of *invected* transcripts may indicate that transcription of *invected* is activated at a more advanced developmental stage, it is also possible that its delay in appearance is a consequence of its large transcription unit: Since the transcription unit of *engrailed* (approximately 4 kb) is considerably smaller than that of *invected* (greater than 30 kb), the transit time for RNA polymerase (estimated to be approximately 1 kb/min; C. Thummel and D. Hogness, pers. comm.) to transcribe *invected* might alone account for the approximately 30-min delay in the appearance of the striped *invected* pattern. Therefore, it is possible that expression of both genes in the embryonic posterior compartments might be activated during the same developmental stage.

The presence of neighboring *Drosophila* genes that share characteristics of expression, sequence, and function so extensively is, except for repeated gene families, rare. Why the apparent duplication, especially in a genome otherwise quite parsimonious? We address this issue first by considering the likelihood that the *invected* gene has a function, and second by considering the implications of the apparent duplication. There are at present no genetic data to indicate whether the *invected* gene is essential, or whether its removal would affect the development of the fly. It is possible that the *invected* RNA is not translated or that inactivation of the gene would be without effect. However, in the shared protein-coding regions, the two genes show more amino acid than nucleotide homology. Moreover, these regions are split by intervening sequences that do not share homology. Thus, at least these regions of the *invected* gene are probably translated. Supporting this is the observation that two genes with sequence homology with *engrailed* and *invected* are present in the mammalian genome (Joyner et al. 1985; Joyner and Martin, this issue). These mammalian genes, *En-1* and *En-2*, are more than 70% homologous over a 107-amino-acid carboxy-terminal region. Thus, if sequence conservation is indicative of essential function, we may predict that the *Drosophila invected* gene is indeed important for development, either of the cells of the posterior developmental

compartments, or of the embryonic nervous system, hindgut, or clypeolabrum. *invected* mutants might therefore be expected to affect embryonic segmentation and/or viability.

The existence of both the *engrailed* and *invected* genes strongly suggests that the homologous portions of the encoded proteins, if carrying out identical biochemical functions, do so in different ways or are regulated differently by those portions of the molecules that are not conserved. It might be argued that the very existence of two such similar genes as *engrailed* and *invected* in fact emphasizes the importance of these nonconserved regions of the proteins whose function or structure is not known. Thus, although the presence of a homeo box in its sequence suggests that *invected*, like *engrailed*, helps to regulate segmental organization during *Drosophila* development, the developmental functions of the two genes may differ despite striking similarities in the structure and biochemical action of the proteins they encode.

Experimental procedures

General procedures

The cDNA library in λgt10 which was described previously (Poole et al. 1985) was screened with a 3.8-kb genomic *Eco*RI fragment from phage E4 (Kuner et al. 1985). The sequence of both strands of the two cDNA clones was determined by the method of Sanger et al. (1977). Subclones for sequencing were obtained after DNase I digestion (Hong 1982) or by subcloning restriction fragments in either M13mp9/10 or pEMBL10/11 (Dente et al. 1983). Genomic sequences were obtained by subcloning DNA fragments from a recombinant library of *Drosophila* DNA derived from Canton S flies (Maniatis et al. 1982). Genomic regions corresponding to the entire *invected* mature transcript, with the exception of the 6-bp Exon II, were sequenced. Intron III and its splice junctions were also sequenced.

Probes were prepared by nick-translation of gel-purified restriction fragments. Single-stranded probes were prepared by extending synthetic primers that were annealed to single-stranded M13 DNA containing appropriate recombinant inserts (Hu and Messing 1982). Northern analyses were as described (Poole et al. 1985).

S1 nuclease mapping

A 539-base genomic subclone (nucleotides −352 to +187; see Fig. 2) in M13mp18 was used to synthesize a uniformly labeled, single-stranded copy of the *invected* noncoding strand (Hu and Messing 1982). Poly(A)$^+$ RNA was prepared as previously described (Poole et al. 1985). Approximately 1×10^5 cpm of the single-stranded probe were combined with 0- to 12-hr embryonic poly(A)$^+$ RNA (10 μg) in S1 hybridization buffer [80% deionized formamide; 0.4 M NaCl; 0.04 M PIPES (pH 6.4); 0.001 M EDTA] at 80°C for 10 min. The mixture was cooled to 52°C; after 4 hr, it was placed on ice, and 200 μl of ice-cold S1 digestion buffer [0.25 M NaCl, 0.3 M $NaCO_2CH_3$(pH 4.6), 0.001 M $ZnSO_4$, 20 μg/ml calf thymus DNA] containing either 26.5, 53, or 106 units of S1 nuclease (Pharmacia) were added. Incubation was for 2 hr at 20°C. The reaction products were extracted with phenol, precipitated with EtOH, resuspended in 90% formamide, and denatured at 68°C immediately before electrophoretic fractionation on a 6% polyacrylamide–8 M urea gel.

Primer extension analysis

Synthetic primers 1 and 2 (0.02 O.D. units)(see Results) were end-labeled with T4 polynucleotide kinase (P-L Biochemicals) in the presence of [γ-^{32}P]dATP (7000 mCi/mmole). After purification on a Sep-Pak column (Waters Associates), 2×10^5 to 4×10^5 Cerenkov cpm of probe were combined in 6 μl H_20 with 10 μg of poly(A)$^+$ RNA that had been isolated from 0- to 12-hr postfertilization embryos. The samples were incubated at 80°C for 2 min and 42°C for 30 min. Then 2 μl of 5× RT buffer [250 mM Tris (pH 8.4), 62.5 mM KCl, 5 mM MgCl, 500 μM dNTP], 1 μl 500 mg/ml Actinomycin D, 0.5 μl of 35 U/μl RNasin (Promega Biotech) containing 16.6 mM DTT, and 0.25 μl reverse transcriptase (200 U/μl, Bethesda Research Labs) were added and the reactions were incubated at 42°C for 45 min. Ethanol-precipitable material was collected and resuspended in 15 μl 95% formamide, 5 μl of which were denatured at 65°C for 10 min, and was fractionated electrophoretically on a 6% acrylamide–8 M urea gel.

The *Eco*RI–*Dde*I fragment (1.7 μg; nucleotides 1–253, see Fig. 2) was also prepared for primer extension analysis. The fragment was treated with calf intestinal alkaline phosphatase (Boehringer Mannheim) and the phosphatase was removed by extraction with phenol. The DNA was extracted 2× with ether, EtOH precipitated, and resuspended in 10 mM Tris (pH 8.0), 1 mM EDTA. The fragment was digested with *Hin*fI and the resulting 240-bp *Hin*F–*Dde*I fragment (positions 13–253; see Fig. 2) was gel purified. The fragment was labeled and treated as above except that approximately 5×10^5 Cerenkov cpm were used.

In situ hybridization

In situ hybridization was performed as described previously (Kornberg et al. 1985; Edgar et al. 1986). [^{35}S]dATP-labeled probes that contained only sequences unique to either *invected* or *engrailed* were prepared by nick-translation of gel-purified restriction fragments (a 1314-bp *Eco*RI–*Xho*I fragment of *invected* cDNA 2.3, positions 28–1342, see Fig. 2; and a 1214-bp *Eco*RI–*Xho*I fragment of *engrailed* cDNA c-2.4, positions 1–1214, see Fig. 2 of Poole et al. 1985). Autoradiography was for 17–27 days at 4°C.

Acknowledgments

We gratefully acknowledge helpful comments on the manuscript from Gail Martin and gifts of imaginal discs from Odessa Eugene and James Fristrom. This work was supported by National Institutes of Health grants to T.K. and postdoctoral fellowships to K.G.C. and W.C.S. S.J.P. and M.P.W. were supported by postdoctoral fellowships from the Weingart Foundation and Jane Coffin Childs Memorial Fund for Medical Research, respectively.

References

Brower, D., P. Lawrence and M. Wilcox. 1981. Clonal analysis of the undifferentiated wing disc of *Drosophila*. *Dev. Biol.* **86:** 448–455.

Bryant, P. 1975. Pattern formation in the imaginal wing disc of *Drosophila melanogaster*: Fate map, regeneration, and duplication. *J. Exp. Zool.* **173:** 49–78.

Dente, L., G. Cesareni, and R. Cortese. 1983. pEMBL: A new family of single stranded plasmids. *Nucleic Acids Res.* **11:** 1645–1655.

DiNardo, S., J.M. Kuner, J. Theis, and P.H. O'Farrell. 1985. Development of embryonic pattern in *D. melanogaster* as re-

vealed by accumulation of the nuclear *engrailed* protein. *Cell* **43:** 59–69.

Edgar, B.A., M.P. Weir, G. Schubiger, and T. Kornberg. 1986. Repression and turnover pattern *fushi tarazu* RNA in the early *Drosophila* embryo. *Cell* **47:** 747–754.

Hong, G. 1982. A systematic DNA sequencing strategy. *J. Mol. Biol.* **158:** 539–549.

Hu, N. and J. Messing. 1982. The making of strand-specific M13 probes. *Gene* **17:** 271–277.

Joyner, A.L., T. Kornberg, K.G. Coleman, D.R. Cox, and G.R. Martin. 1985. Expression during embryogenesis of a mouse gene with sequence homology to the *Drosophila engrailed* gene. *Cell* **43:** 29–37.

Kaufman, T.C., R. Lewis, and B. Wakimoto. 1980. Cytogenetic analysis of chromosome 3 in *Drosophila melanogaster*: The homeotic gene complex in polytene chromosome interval 84A-B. *Genetics* **94:** 115–133.

Kornberg, T. 1981a. *engrailed*: A gene controlling compartment and segment formation in *Drosophila*. *Proc. Natl. Acad. Sci.* **78:** 1095–1099.

———.1981b. Compartments in the abdomen of *Drosophila* and the role of the *engrailed* locus. *Dev. Biol.* **86**: 363–372.

Kornberg, T., I. Siden, P. O'Farrell, and M. Simon. 1985. The *engrailed* locus of *Drosophila*: *in situ* localization of transcripts reveals compartment-specific expression. *Cell* **40:** 45–53.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* **12:** 857–872.

Kuner, J.M., M. Nakanishi, Z. Ali, B. Drees, E. Gustavson, J. Theis, L. Kauvar, T. Kornberg, and P.H. O'Farrell. 1985. Molecular cloning of *engrailed*: A gene involved in the development of pattern in *Drosophila melanogaster*. *Cell* **42:** 309–316.

Lawrence, P.A. and G. Morata. 1976. Compartments in the wing of *Drosophila*: A study of the *engrailed* gene. *Dev. Biol.* **50:** 321–337.

Lawrence, P.A. and G. Struhl. 1982. Further studies on the *engrailed* phenotype in *Drosophila*. *EMBO J.* **1:** 827–833.

Lewis, E. 1967. Genes and gene complexes. In *Heritage from Mendel* (ed. R.A. Brink and E.D. Styles), pp. 17–47. University of Wisconsin Press, Madison.

———.1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276:** 565–570.

Maniatis, T., E.F. Fritsch, and J. Sambrook. 1982. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

McGinnis, W., R.L. Garber, J. Wirz, A. Kuroiwa, and W.J. Gehring. 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37:** 403–408.

Poole, S.J., L.M. Kauvar, B. Drees, and T. Kornberg. 1985. The *engrailed* locus of *Drosophila*: Structural analysis of an embryonic transcript. *Cell* **40:** 37–43.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Weir, M.P. and T. Kornberg. 1985. Patterns of *engrailed* and *fushi tarazu* transcripts reveal novel intermediate stages in *Drosophila* segmentation. *Nature* **318:**433–445.