Development 114, 221-232 (1992)
Printed in Great Britain © The Company of Biologists Limited 1992

The *Drosophila pumilio* gene: an unusually long transcription unit and an unusual protein

PAUL M. MACDONALD

Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA

Summary

Specification of the posterior body plan in *Drosophila* requires the action of a determinant prelocalized to the posterior pole of the embryo. During embryogenesis this determinant appears to move anteriorly in a process dependent on the *pumilio* (*pum*) gene. This report describes the cloning and molecular characterization of a cDNA derived from the *pum* gene, and the analysis of *pum* mRNA and protein expression during early *Drosophila* development. The *pum* gene is unusually large; comparison of genomic and cDNA sequences reveals that the pum transcription unit is at least 160 kb in length. The *pum* cDNA encodes a 157×10^3 M_r protein which consists mainly of regions enriched in a single amino acid, usually glycine, alanine, glutamine or serine/threonine. Six tandem repeats of a 36 amino acid repeat unit are also present. Pum protein is cytoplasmic and is concentrated in a subcortical region of the embryo. The distribution of pum protein exhibits no asymmetry along the anteroposterior axis of the embryo.

Key words: *pumilio* gene, *Drosophila*, transcription unit, specification, posterior body plan.

Introduction

In *Drosophila*, specification of the global body plan is directed by several localized determinants, or spatial cues. These determinants are provided by the mother and are deposited in the egg during oogenesis. In some cases, a specific molecule distributed along the length of the egg is activated locally to form the determinant and, in other cases, the determinant is itself localized during oogenesis. The systems that specify anterior and posterior pattern elements along the anteroposterior body axis are examples of the latter type.

In the anterior system, the determinant is the product of the *bicoid* (*bcd*) gene (Frohnhöfer and Nüsslein-Volhard, 1986). During oogenesis, *bcd* mRNA is deposited at the anterior pole of the oocyte (Berleth et al., 1988; Frigerio et al., 1986). Following fertilization, *bcd* mRNA becomes translationally active. bcd protein accumulates and diffuses away from the anterior pole of the embryo to form a concentration gradient (Driever and Nüsslein-Volhard, 1988a). This gradient specifies positional information, with high levels of protein directing formation of the most anterior structures and progressively lower values directing formation of progressively more posterior body parts (Driever and Nüsslein-Volhard, 1988b; Driever and Nüsslein-Volhard, 1989; Struhl et al., 1989; Driever et al., 1989). Formation of the bcd protein gradient seems to depend primarily on *bcd* mRNA localization during oogenesis, and no genes have been identified that are required specifically to generate the protein gradient from the localized source of mRNA.

In the posterior patterning system, the determinant is encoded by the *nanos* (*nos*) gene (Nüsslein-Volhard et al., 1987; Sander and Lehmann, 1988). As observed for *bcd*, *nos* is localized during oogenesis, and *nos* activity can be detected only at the posterior pole of the embryo (Lehmann and Nüsslein-Volhard, 1986; Frohnhöfer et al., 1986; Sander and Lehmann, 1988). Posterior localization and activation of *nos* activity requires the action of a group of eight genes (Lehmann and Nüsslein-Volhard, 1986, 1987; Boswell and Mahowald, 1985; Nüsslein-Volhard et al., 1987; Manseau and Schüpbach, 1989). Interestingly, all but one of these genes are also required for posterior localization of germ line determinants, implying the existence of a shared localization pathway. Localization of *nos* to the extreme posterior pole does not appear to be sufficient for its function; although *nos* activity is concentrated at the extreme posterior pole of the embryo, its site of action is somewhat anterior in the presumptive abdominal region. This was demonstrated in an elegant series of cytoplasmic transplantation experiments perfomed by Lehmann and Nüsslein-Volhard (1986). Cytoplasm from a wild-type donor embryo was injected into a recipient embryo lacking one of the genes required for *nos* activity. After a growth period, the cuticle was examined to monitor the degree to which the injected cytoplasm could rescue the mutant phenotype. They recovered the most *nos* activity from the posterior pole of donor embryos, but found that the most efficient rescue occurred when this cytoplasm was injected into the recipient at about 20-40% egg length (0% is the posterior end). Thus, *nos* localization seems to involve two steps: localization to the posterior pole; and transfer to the more anterior site of action.

A gene required for the second phase of localization was identified by Lehmann and Nüsslein-Volhard (1987). They performed transplantation experiments similar to those described above, but with the donor and the recipient having the same genotype. Remarkably, cytoplasm taken from the posterior pole of a *pumilio*$^-$ (*pum*$^-$) donor and injected into the abdominal region of a *pum*$^-$ recipient was able partially to rescue the mutant phenotype. Thus, in *pum*$^-$ embryos *nos* activity must be present at the posterior pole, but is somehow unable to act at its target. This result led to the suggestion that *pum* was involved in the transport of *nos* (Lehmann and Nüsslein-Volhard, 1987). Notably, *pum* is the only gene required for *nos* activity that is not also required for posterior localization of germ line determinants, further supporting the notion that there is a second step in the localization of *nos*.

Here I describe the cloning and molecular characterization of *pum* gene cDNAs. The gene is unusually large, with the cDNAs derived from a genomic region of about 160 kb. The protein encoded by the cDNA is also unusual, in that it consists largely of regions rich in one of a number of single amino acids, respectively. A notable exception is one region made up of six degenerate repeats of a 36 amino acid motif. There is no striking similarity to characterized molecular motor proteins, or to any other known protein sequences, and so the mechanism of action of the pum protein remains uncertain. Expression of *pum* was examined by monitoring the distributions of *pum* mRNA and pum protein during oogenesis and early embryogenesis. Curiously, although regional concentration of *pum* mRNA at the posterior pole of early stage embryos was observed, this localization was not consistent among a population of similarly stages embryos, and therefore may not have functional significance. In addition, pum protein was never found to be similarly localized. Thus, there is as yet no indication that *pum* acts in a spatially restricted fashion.

Materials and methods

Chromosome walking and breakpoint mapping

The *In(3R)Msc* chromosome was used to jump from the *Antp* gene complex into the 85C region. DNA was isolated from *In(3R)Msc/TM3* males, partially digested with Sau3A and used to prepare a genomic phage library in lambda DASH (Stratagene). The library was screened with a labeled DNA restriction fragment from the *ftz* upstream region, where the proximal breakpoint of the inversion was reported to map (Scott et al., 1983; Kuroiwa et al., 1985). None of the phage isolated in the first round of screening carried the inversion, and so a short chromosomal walk was carried out in this library until the breakpoint was reached at approximately coordinate +35 of Scott et al. (1983). Phage containing the inversion breakpoint were purified, and a restriction fragment from the 85C region was used to screen a wild type genomic phage library. Several overlapping phage clones were isolated. Various restriction fragments from the 85C genomic regions flanking the position of the inversion breakpoint were labeled and used to probe an ovarian cDNA library (from

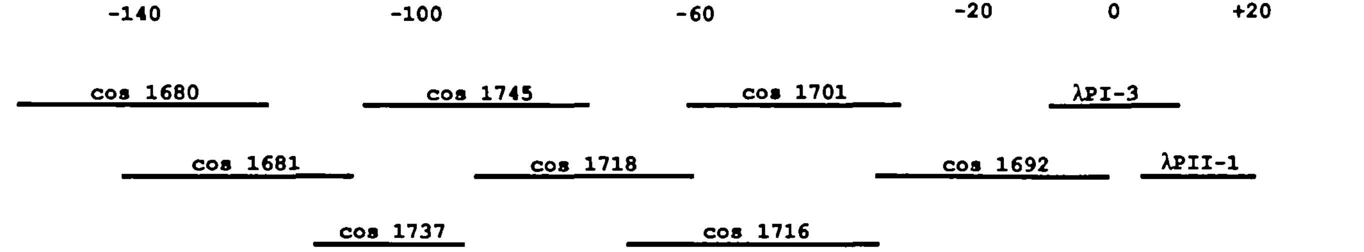


Fig. 1. Cloning of the *pum* gene. Chromosome walk in the *pum* region. The entry point was at the inversion breakpoint of the *In(3R)Msc* chromosome (coordinate 0 on the map; coordinates indicate kb from the inversion breakpoint). The map is oriented with the centromere-proximal region at the left, and the region towards the telomere at the right. Phage and cosmid clones covering the walk are shown below the map. More cosmids were isolated than are shown. Note that cosmid 1737 is smaller than the expected minimum size for packaging in phage lambda capsids. Since there are no other cosmids covering this region, it is possible that a segment of *Drosophila* genomic DNA was deleted during propagation of cosmid 1737 and is therefore missing from the map. The segment of genomic DNA deleted in *Df(3R)p*^{XT9} is shown as a hatched bar; the uncertainty in the exact position of the breakpoint is indicated by an open bar. Approximate positions of exons of the *pum* transcript are indicated by vertical lines (not drawn to scale) and identified by number above the map.

Laura Kalfayan). Some of these fragments contained repetitive elements and so were not useful. Of the fragments that contained only single copy DNA, several hybridized in independent screenings to phage in the cDNA library, and cDNA clones from each screen were purified. All are apparently derived from the same mRNA species; they crosshybridized to one another, their restriction maps were superimposable in the regions of overlap, and they all detected an approximately 7 kb mRNA in northern blot analysis of early embryo RNA. In order to clone the remainder of the genomic DNA corresponding to the cDNA, hybridization probes were prepared from parts of the cDNA and used to screen genomic libraries. Phage libraries were first screened. When it became apparent that the gene might be very large, a cosmid library was obtained (from J. Tamkun and M. Scott) for further screenings. Gaps between the initial genomic clones were filled by chromosomal walking in the cosmid library. After the walk was complete, the positions of the distal breakpoints in *Df(3R)p*^{XT103} and *Df(3R)p*^{XT9} were determined by Southern blot analysis using DNA from flies heterozygous for one of the deficiency chromosomes and for the TM3 Sb balancer chromosome.

cDNA and transcript characterization

Two cDNA clones that together cover the largest genomic region were chosen for more detailed analysis and subcloned into plasmid vectors, creating p1617 and p1619. A general outline of intron/exon structure was derived by probing Southern blots of the entire walk with various parts of the cDNAs. The direction of transcription was determined by RNAase protection mapping (Zinn et al., 1983). The DNA sequence of the entire cDNA clone in p1619 was obtained following construction of two nested sets of deletion mutants, one extending from either end of the clone. The most 5′ portion of this cDNA was different from all other *pum* cDNAs, and was probably added to a partially reverse-transcribed *pum* cDNA during cloning. Although it would be difficult to prove that the two parts were not spliced together in vivo, I have been unable to detect any such RNA in RNAase protection experiments using a probe covering the junction of the two parts. To determine more of the *pum* cDNA 5′ sequence shared by all other *pum* cDNAs isolated, the 5′-most region of the cDNA in p1617 was sequenced after subcloning restriction fragments into pEMBL8 or pEMBL9 (Dente et al., 1983). In addition a further cDNA was isolated by reprobing the library with a labeled *SmaI-XhoI* restriction fragment (nucleotides 1260-1497 in Fig. 3) to obtain more 5′ sequences. The 5′ regions of this cDNA was also sequenced. Various genomic DNA fragments were also cloned into pEMBL8 or pEMBL9 for sequencing. DNA sequencing was initially performed with Klenow fragment, and later with Sequenase version 2.0 (US Biochemicals). Much of the *pum* cDNA was GC-rich and compressions were frequently observed in DNA sequencing ladders. To obtain accurate sequence data, both strands of the cDNA were sequenced throughout, and all reactions were perfomed in duplicate using both dGTP and dITP. Genomic and cDNA sequences were compared to determine the exact intron/exon organization.

Antibodies and immunological methods

Two different clones were constructed to allow expression of parts of pum protein in *E. coli*. In p1637, a *XhoI-EcoRI* fragment (nucleotide 1497 to a linker at the end of the sequence in Fig. 3) was first modified by addition of a *BglII* linker at the blunted *XhoI* site, and then inserted into the *BamHI* and *EcoRI* sites of pET3b (Rosenberg et al., 1987). The protein expressed from this construct, referred to as pum polypeptide A, includes amino acids 392-1533 of *pum* fused at the N terminus to 11 amino acids from phage T7 gene *10* and the amino acids encoded by the linker sequence. In p1648, a *BamHI-EcoRI* fragment (nucleotides 3424 to a linker at the end of the sequence in Fig. 3) was inserted into the *BamHI* and *EcoRI* sites of pET3c. The protein expressed from this construct, refered to as pum polypeptide B, includes amino acids 1034-1533 of *pum* fused at the N terminus to 11 amino acids from phage T7 gene *10* and the amino acids encoded by the linker sequence. Both proteins were partially purified as

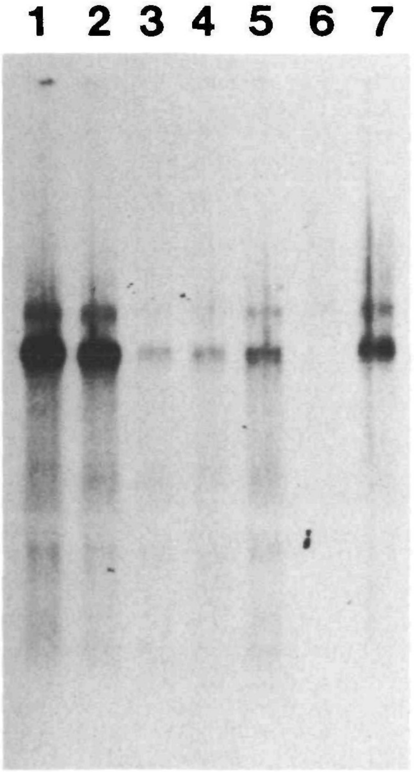


Fig. 2. Northern blot analysis of *pum* mRNA. A northern blot was prepared by electrophoresing 10 μg of total RNA from various developmental stages in a 1.0% agarose gel containing formaldehyde and transferring the RNA to GeneScreen membrane. The blot was probed with a genomic *ClaI-SacI* restriction fragment (nucleotides approx. −500 to 112 in Fig. 3) containing sequences from the first exon (similar results have been obtained using several other probes lacking repetitive sequences:
a *SmaI-XhoI* restriction fragment, nucleotides 1260-1497;
a *PstI* restriction fragment, nucleotides 1528-1958;
a *BglII-MluI* restriction fragment, nucleotides 3598-4433).
Sources of RNA samples were: lane 1, 0-3 hour embryos; lane 2, 3-6 hour embryos; lane 3, 6-12 hour embryos; lane 4, 12-24 hour embryos; lane 5, pupae; lane 6, adult males; lane 7, adult females. Two transcripts are detected: an approximately 7.0 kb mRNA is present at highest levels in early embryos and adult females, at lower levels in later embryonic stages and in pupae, and is not detectable in adult males; a larger and rarer mRNA is present in all developmental stages examined, being most abundant in early embryos, pupae and adult females, and less abundant in late embryos and adult males.

```
AGTGTTGCAAAACGCGCGTGTGGTTCCTTGTGCTGCAAGTTAAAATACAATTCAAGTTGGCAATACGCGCAAAATTGTCAGCTGCGATAGCTAGGAAAAGCCTCCAAAATTGAGCTCCTA    120

ACCGCGCCCACAATTGCCATATCGACGCCCTCGCCGCAGCAGCAACACCAACAGCAGCAGCAGCAGCAGCAGCAACTCTATCAGCAACATCAACAGCAGCAGCAGCAACATTACGGT    240

CCACCACCGCCCTACTTTCAACAGCTACACCAGCAACACCAACAGCAGCAGCAACAACAGCAGCAGCAGCAACACCAGCAACACATGAAGTTTTTGGGTGGTAACGATGATCGCAATGGC    360
                                                                                    M  K  F  L  G  G  N  D  D  R  N  G        12

CGCGGAGGCGTCGGCGTTGGCACGGATGCCATTGTAGGATCTCGAGGTGGCGTCTCTCAGGATGCCGCCGATGCAGCTGGTGCCGCCGCAGCCGCCGCCGTCGGCTATGTCTTCCAGCAG    480
R  G  G  V  G  V  G  T  D  A  I  V  G  S  R  G  G  V  S  Q  D  A  A  D  A  A  G  A  A  A  A  A  A  V  G  Y  V  F  Q  Q        52

CGTCCATCGCCTGGTGGGGTTGGCGTCGGCGTGGGCGGAGTGGGTGGCGGTGTGCCAGGGGTCGGAGCCGTAGGCTCGACCTTGCACGAGGCCGCCGCCGCCGAGTACGCCGCCCACTTT    600
R  P  S  P  G  G  V  G  V  G  V  G  G  V  G  G  G  V  P  G  V  G  A  V  G  S  T  L  H  E  A  A  A  A  E  Y  A  A  H  F        92
                                                         12
GCCCAGAAGCAACAGCAGACCCGATGGGCGTGCGGCGACGACGGCCATGGGATCGATAACCCGGACAAATGGAAGTACAATCCGCCGATGAATCCGGCCAATGCCGCTCCTGGCGGTCCA    720
A  Q  K  Q  Q  Q  T  R  W  A  C  G  D  D  G  H  G  I  D  N  P  D  K  W  K  Y  N  P  P  M  N  P  A  N  A  A  P  G  G  P       132

CCGGGAAATGGCAGTAATGGTGGGCCCGGCGCCATTGGAACCATTGGCATGGGCAGCGGATTGGGTGGTGGTGGCGGCGGCGGAGCTGGCGGCGGAAATAATGGCGGCTCTGGTACGAAT    840
P  G  N  G  S  N  G  G  P  G  A  I  G  T  I  G  M  G  S  G  L  G  G  G  G  G  G  G  A  G  G  G  N  N  G  G  S  G  T  N       172

GGCGGTCTGCATCATCAATCGATGGCCGCTGCAGCTGCGAATATGGCAGCCATGCAACAGGCGGCGGCGTTGGCCAAGCACAATCACATGATATCACAGGCAGCAGCCGCAGTTGCAGCC    960
G  G  L  H  H  Q  S  M  A  A  A  A  A  N  M  A  A  M  Q  Q  A  A  A  L  A  K  H  N  H  M  I  S  Q  A  A  A  A  V  A  A       212

CAGCAACAACATCAGCATCCACACCAGCAGCATCCCCAGCAGCAGCAGCAACAGCAGCAGGCGCAGAACCAGGGGCATCCACATCACCTTATGGGCGGTGGCAATGGACTGGGCAACGGC   1080
Q  Q  Q  H  Q  H  P  H  Q  Q  H  P  Q  Q  Q  Q  Q  Q  Q  Q  A  Q  N  Q  G  H  P  H  H  L  M  G  G  G  N  G  L  G  N  G       252

AATGGATTGGGCATACAACATCCCGGCCAGCAACAGCAGCAGCAGCAGCAACAACAGCAGCAGCAACATCCCGGCCAGTACAACGCGAATCTGCTTAACCATGCGGCTGCCTTGGGTCAC   1200
N  G  L  G  I  Q  H  P  G  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  H  P  G  Q  Y  N  A  N  L  L  N  H  A  A  A  L  G  H       292

ATGTCATCTTATGCCCAATCGGGTGGCAGCATGTACGACCATCATGGTGGAGCCATGCACCCGGGAATGAACGGCGGCATGCCCAAGCAACAGCCATTGGGTCCACCCGGAGCCGGAGGA   1320
M  S  S  Y  A  Q  S  G  G  S  M  Y  D  H  H  G  G  A  M  H  P  G  M  N  G  G  M  P  K  Q  Q  P  L  G  P  P  G  A  G  G       332
                                                                                                                       2
CCCCAGGACTATGTCTACATGGGTGGCCAGACCACTGTGCCCATGGGAGCCGCAATGATGCCGCCACAGAATCAATATATGAACAGCGCTGCTGTTGCAGCTGCCAATCGGAATGCAGCG   1440
P  Q  D  Y  V  Y  M  G  G  Q  T  T  V  P  M  G  A  A  M  M  P  P  Q  N  Q  Y  M  N  S  A  A  V  A  A  A  N  R  N  A  A       372
3
ATTACCACATCCACTGCCAAGAAATTGTGGGAGAAATCCGATGGCAAGGGCGTATCCTCGAGCACTCCCGGTGGACCGTTGCATCCCCTGCAGATCCCCGGCATCGGGGATCCCTCCTCC   1560
I  T  T  S  T  A  K  K  L  W  E  K  S  D  G  K  G  V  S  S  S  T  P  G  G  P  L  H  P  L  Q  I  P  G  I  G  D  P  S  S       412

GTGTGGAAGGATCACACCTGGTCCACACAGGGCGAGAATATATTGGTGCCGCCCCCCTCGCGAGCCTACGCCCATGGAGGCGCCTCCGATACTTCAAACAGCGGCAATGCGGGCATACTG   1680
V  W  K  D  H  T  W  S  T  Q  G  E  N  I  L  V  P  P  P  S  R  A  Y  A  H  G  G  A  S  D  T  S  N  S  G  N  A  G  I  L       452
 34
AGTCCCCGCGATTCGACTTGCGCCAAAGTGGTTGAATATGTTTTCAGTGGCTCGCCCACCAACAAAGATAGCTCGCTTTCCGGATTGGAACCGCATTTGCGGAATCTAAAGTTTGACGAC   1800
S  P  R  D  S  T  C  A  K  V  V  E  Y  V  F  S  G  S  P  T  N  K  D  S  S  L  S  G  L  E  P  H  L  R  N  L  K  F  D  D       492
          45
AACGATAAGTÇACGCGACGATAAGGAGAAAGCAAACTCTCCGTTTGACACAAACGGTTTGAAGAAAGACGATCAGGTCACAAACTCAAATGGTGTTGTCAACGGCATTGACGATGACAAG   1920
N  D  K  S  R  D  D  K  E  K  A  N  S  P  F  D  T  N  G  L  K  K  D  D  Q  V  T  N  S  N  G  V  V  N  G  I  D  D  D  K       532
        56
GGCTTCAATCGCACTCCTGGTTCACGTCAACCATCACCTGCAGAGGAGTCCCAGCCACGTCCCCCCAATCTACTCTTTCCTCCACTGCCCTTCAATCACATGCTCATGGATCATGGCCAA   2040
G  F  N  R  T  P  G  S  R  Q  P  S  P  A  E  E  S  Q  P  R  P  P  N  L  L  F  P  P  L  P  F  N  H  M  L  M  D  H  G  Q       572

GGCATGGGAGGCGGCTTGGGCGGAGTTGTTGGATCTGGCAACGGAGTCGGCGGTGGCAGCGGCGGAGGCGGGGCAGGCGGCGCTTATGCGGCCCACCAGCAGATGGCCGCCCAGATGAGT   2160
G  M  G  G  G  L  G  G  V  V  G  S  G  N  G  V  G  G  G  S  G  G  G  G  A  G  G  A  Y  A  A  H  Q  Q  M  A  A  Q  M  S       612
                                                         67
CAATTGCAACCGCCGATGATGAACGGCGTTGGCGGCGGAATGCCAATGGCAGCACAGTCACCAATGTTGAATCACCAGGCAGCTGGACCCAATCACATGGAATCTCCCGGAAATCTCTTG   2280
Q  L  Q  P  P  M  M  N  G  V  G  G  G  M  P  M  A  A  Q  S  P  M  L  N  H  Q  A  A  G  P  N  H  M  E  S  P  G  N  L  L       652
                        78
CAGCAGCAAAATTTTGATGTTCAGCAACTGTTTCGCTCGCAGAATCCGGGCCTAGCAGCAGTTGCCACAAATGCAGCGGCCGCAGCAGCAGCCGCAGCAGCTGCCACATCGGCAGCGAGT   2400
Q  Q  Q  N  F  D  V  Q  Q  L  F  R  S  Q  N  P  G  L  A  A  V  A  T  N  A  A  A  A  A  A  A  A  A  A  A  T  S  A  A  S       692

GCTGCGGCAGCGGTGGGCGCACCACCCGTTCCCAACGGATCGCTGCAGCAGTCGCAGCAGCAACAGCAGCAGCAGCAACAACAGCAGCAGCAACAACAGATGCACATGGCGGCCGCGTCG   2520
A  A  A  A  V  G  A  P  P  V  P  N  G  S  L  Q  Q  S  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  M  H  M  A  A  A  S       732

CAACAATTTTTGGCCGCCCAGCAGCAGGCGCAAAATGCGGCCTATGCCGCCCAACAGGCCACGTCCTACGTCATCAATCCGGGCCAGGAGGCTGCCCCGTATATGGGCATGATTGCCGCC   2640
Q  Q  F  L  A  A  Q  Q  Q  A  Q  N  A  A  Y  A  A  Q  Q  A  T  S  Y  V  I  N  P  G  Q  E  A  A  P  Y  M  G  M  I  A  A       772

GCCCAGATGCCGTACTATGGCGTAGCACCATGGGGCATGTATCCGGGCAATCTGATTCCGCAACAGGGAACGCAGCCGCGCCGCCCCCTCACCCCCTCGCAGCAGGGTGCCGAGAATCAG   2760
A  Q  M  P  Y  Y  G  V  A  P  W  G  M  Y  P  G  N  L  I  P  Q  Q  G  T  Q  P  R  R  P  L  T  P  S  Q  Q  G  A  E  N  Q       812
          89
CCGTATCAGGTCATCCCGGCATTCCTCGATCACACGGGCTCCTTGCTGATGGGAGGACCTCGCACCGGGACGCCGATGCGTCTGGTTAGCCCCGCCCCCGTTCTGGTGCCCCCGGGCGCT   2880
P  Y  Q  V  I  P  A  F  L  D  H  T  G  S  L  L  M  G  G  P  R  T  G  T  P  M  R  L  V  S  P  A  P  V  L  V  P  P  G  A       852
                                                                                                          910
ACCCGTGCCGGCCCCCCGCCCCCGCAGGGCCCACAGCTGTATCAGCCGCAGCCGCAGACGGCCCAACAGAATCTCTACTCGCAGCAGAATGGATCCAGTGTCGGAGGCCTCGCCTTGAAC   3000
T  R  A  G  P  P  P  P  Q  G  P  Q  L  Y  Q  P  Q  P  Q  T  A  Q  Q  N  L  Y  S  Q  Q  N  G  S  S  V  G  G  L  A  L  N       892

ACGAGCTCGTTGACGGGTCGCCGCGACTCCTTCGACCGCAGCACCTCCGCCTTCAGTCCCTCGACCATGGACTACACCAGCAGCGGTGTGGCAGCGGCCGCCAATGCGGTGAACAGCACA   3120
T  S  S  L  T  G  R  R  D  S  F  D  R  S  T  S  A  F  S  P  S  T  M  D  Y  T  S  S  G  V  A  A  A  A  N  A  V  N  S  T       932

GTGGCCCAGGCAGCAGCAGCTGCCGCAGCAGCCGCCGCAGCGCGTGGCAAGTGGCCGGGAGCGATGTCGGGAGCGGCCAGTGGAGCCTACGGAGCCCTGGGAGCGGGCAATGCTTCGGCC   3240
V  A  Q  A  A  A  A  A  A  A  A  A  A  A  R  G  K  W  P  G  A  M  S  G  A  A  S  G  A  Y  G  A  L  G  A  G  N  A  S  A       972

AGTCCCCTGGGCGCACCAATCACGCCGCCGCCATCGGCGCAATCCTGTCTCCTGGGCAGTCGGGCACCTGGAGCCGAGTCCCGCCAGCGGCAGCAGCAACAACAGCAGCTGGCCGCCGTT   3360
S  P  L  G  A  P  I  T  P  P  P  S  A  Q  S  C  L  L  G  S  R  A  P  G  A  E  S  R  Q  R  Q  Q  Q  Q  Q  Q  L  A  A  V      1012

GGTCTGCCGGCGACTGCAGCAGCTGCTCAGGCAGCGGTGGCCGCGGCTGCCAACAATATGTTCGGATCCAACAGCTCGATCTTCTCGAATCCCCTGGCCATTCCGGGTACCGCAGCTGTG   3480
G  L  P  A  T  A  A  A  A  Q  A  A  V  A  A  A  A  N  N  M  F  G  S  N  S  S  I  F  S  N  P  L  A  I  P  G  T  A  A  V      1052

GCAGCTGCAGCGGCAGCAGCAGCGGCCGCCAACTCGCGTCAGGTGGCTGCCACGGCAGCGGCAGCAGCGGCGGTGGCAGCAGCAGCCGGCGGAGTGGGAGGTGCCCCACAGCCAGGAAGA   3600
A  A  A  A  A  A  A  A  A  A  N  S  R  Q  V  A  A  T  A  A  A  A  A  A  V  A  A  A  A  G  G  V  G  G  A  P  Q  P  G  R      1092

TCTCGCCTTCTCGAAGATTTCCGCAACCAGCGGTATCCAAATCTTCAGCTACGCGATCTCGCTAACCACATTGTGGAGTTCTCACAGGATCAGCACGGCTCGCGGTTTATCCAACAGAAG   3720
S  R  L  L  E  D  F  R  N  Q  R  Y  P  N  L  Q  L  R  D  L  A  N  H  I  V  E  F  S  Q  D  Q  H  G  S  R  F  I  Q  Q  K      1132
```

```
TTGGAGCGGGCCACCGCCGCCGAGAAGCAAATGGTGTTCAGCGAGATCCTGGCGGCAGCCTATAGCCTGATGACCGATGTCTTTGGCAACTATGTCATCCAGAAGTTCTTTGAGTTCGGC   3840
L  E  R  A  T  A  A  E  K  Q  M  V  F  S  E  I  L  A  A  A  Y  S  L  M  T  D  V  F  G  N  Y  V  I  Q  K  F  F  E  F  G     1172

ACTCCCGAGCAGAAGAACACGCTGGGCATGCAGGTCAAGGGTCATGTGCTGCAGCTGGCGCTGCAAATGTATGGCTGCCGAGTGATTCAGAAGGCTCTGGAGAGCATCTCGCCGGAGCAG   3960
T  P  E  Q  K  N  T  L  G  M  Q  V  K  G  H  V  L  Q  L  A  L  Q  M  Y  G  C  R  V  I  Q  K  A  L  E  S  I  S  P  E  Q     1212

CAGCAGGAAATCGTCCACGAACTGGACGGACATGTGCTGAAATGCGTCAAGGATCAGAATGGCAATCATGTGGTGCAAAAGTGCATTGAGTGCGTGGACCCCGTGGCGCTGCAGTTCATC   4080
Q  Q  E  I  V  H  E  L  D  G  H  V  L  K  C  V  K  D  Q  N  G  N  H  V  V  Q  K  C  I  E  C  V  D  P  V  A  L  Q  F  I     1252

ATCAATGCTTTCAAGGGTCAGGTTTACTCGCTAAGCACCCATCCGTATGGATGCCGGGTGATCCAGAGAATCCTTGAGCATTGCACTGCCGAACAGACCACGCCCATTTTGGACGAACTG   4200
I  N  A  F  K  G  Q  V  Y  S  L  S  T  H  P  Y  G  C  R  V  I  Q  R  I  L  E  H  C  T  A  E  Q  T  T  P  I  L  D  E  L     1292
                                                               1011
CATGAGCACACCGAACAGTTGATTCAGGACCAATATGGCAACTATGTTATTCAGCATGTGCTTGAACACGGCAAGCAGGAGGATAAGTCGATTCTTATCAACAGCGTGCGCGGCAAAGTT   4320
H  E  H  T  E  Q  L  I  Q  D  Q  Y  G  N  Y  V  I  Q  H  V  L  E  H  G  K  Q  E  D  K  S  I  L  I  N  S  V  R  G  K  V     1332
                                                                                                                1112
CTGGTGCTATCACAGCACAAGTTCGCCTCAAACGTTGTGGAGAAATGTGTTACCCATGCCACTCGCGGAGAACGCACTGGTCTCATAGACGAGGTCTGCACCTTCAACGACAACGCGTTG   4440
L  V  L  S  Q  H  K  F  A  S  N  V  V  E  K  C  V  T  H  A  T  R  G  E  R  T  G  L  I  D  E  V  C  T  F  N  D  N  A  L     1372

CACGTGATGATGAAGGATCAGTATGCCAACTATGTGGTCCAAAAAATGATCGATGTATCGGAGCCGACGCAGCTCAAGAAGCTGATGACCAAGATCCGGAAAAACATGGCCGCCTTGCGC   4560
H  V  M  M  K  D  Q  Y  A  N  Y  V  V  Q  K  M  I  D  V  S  E  P  T  Q  L  K  K  L  M  T  K  I  R  P  H  M  A  A  L  R     1412

AAGTACACCTACGGCAAGCACATCAATGCCAAGTTGGAGAAGTACTACATGAAGATAACCAATCCCATTACGGTGGGCACAGGAGCTGGAGGAGTGCCGGCAGCCTCGTCGGCGGCCGCA   4680
K  Y  T  Y  G  K  H  I  N  A  K  L  E  K  Y  Y  M  K  I  T  N  P  I  T  V  G  T  G  A  G  G  V  P  A  A  S  S  A  A  A     1452

GTCAGCAGTGGTGCCACCTCGGCATCGGTAACCGCCTGCACCAGTGGCAGCAGCACCACCACGACCAGCACTACCAACAGCCTGGCCTCACCCACCATTTGTTCGGTGCAGGAGAACGGC   4800
V  S  S  G  A  T  S  A  S  V  T  A  C  T  S  G  S  S  T  T  T  T  S  T  T  N  S  L  A  S  P  T  I  C  S  V  Q  E  N  G     1492

AGCGCCATGGTTGTGGAGCCCTCCTCCCCGGACGCCTCCGAGTCCTCGTCCTCGGTGGTGTCAGGCGCTGTCAACAGCAGCTTGGGTCCCATTGGACCCCCGACCAACGGCAACGTTGTG   4920
S  A  M  V  V  E  P  S  S  P  D  A  S  E  S  S  S  S  V  V  S  G  A  V  N  S  S  L  G  P  I  G  P  P  T  N  G  N  V  V     1532

CTGTAAAGGAAATAACAAATTAAGCCAGGCAGTCAAAGGAAACTTCCTTCTCGAATCGCAGTATAGTTTTTAGAAGCTGTAGAGCTTAACATAAACAACAAGTACATATAAATGTAATCT   5040
L  Stop                                                                                                                    1533

TATTTATTGGAAAAGCAGCGATAAATGGAGCTGCACTCGAAGATTTGCAAAGAGGATAGTAAAACACACATGCGCCAATCTAGAGAAACAAATAGCAAACAAAGAAGCACACTGGCAAGC   5160

AAAAAAGCAAAAGAGCTTAACAGCTAAAACTAAAAGAAATTTGTATTTTTACGAACAAAACTAATAACGTTCTCATGAAAAAAGATTTCAAAATATTTGTAAAATGCGCTCGCATAATTA   5280

ATTTGTAAAAAAAAGGCATGAACCGCAAAGATGAAAGAAAACAAAAATGCGTAGTAAATCGCGATCAAGAAAAAAAAATAATGAATGTAATGTAAAATGTCAATGAAACAGATTTGTCTGC   5400

GTACATTTTCGTTGTAACTTTGTATAAATTAATTATTATATAGCAAGTCTATCTGTAAATGATTAATGTTTCGACTGTAAATTAATAAGAAGACAACTGAAGAGCCGGCGAGCTGAAAAA   5520

AAATAAAGTAAAAAGAGCGGGCTGCATGAATTAGCCTACGATTTATAAGTTCAGACAGAGGAACCATTTCTAATATACAAACATATATACGAGGGATAACAGCAGAAGCCGCACTTAGTG   5640

TAGAATGTAGAGTAATAATGTTTTTGGAGCCAGCAGCTACAAAGACACAATGAAAACAGAGACACACGAGACACGCCCACGCCCCCTCACGCACACTCGGTTGCATACACCCACACAATG   5760

AACGACTCTTCAGCCCATTCACGTTGCTTTTGCACTATGTAAAAATTTTGTATAAAAAAAAACCCCAAACAACAAACCATGTAAACCATGTAATTTTCAAATGTTTCACTGTAAAATGTA   5880

TACATACTTTATTTTGTAAATTTTTTTTAAGTCGCAAGTAACTCATACATATTCTATTCTAAACCTCACGCATGTATTTATAATTTTATACACATTAGCTGGTGACCACCGATCGACGAT   6000

CTGCATGGATGTTGGTCAGCTGGTGGCCAGCTAAAAGAACCTGTTAGCCAAGTAAGCCAAAAATGATAATAATTGGATTTTAAAACAATAACCATCAAAATAAACCAATTTTTTTCAAAA   6120
```

Fig. 3. DNA sequence of the *pum* cDNA. The sequence was assembled from three cDNAs. These cDNAs do not extend to the 5′ end of the transcript, as shown by RNAase protection assays using probes derived from the adjacent genomic DNA region (data not shown). Nevertheless, the single long ORF in the cDNA is fully contained within the region shown, since it is blocked upstream in the cDNA by an in-frame stop codon. The sequence ends at the start of a run of A residues in the cDNA clone that are not present in the genomic sequence. This is likely to be the polyadenylation site, since it is preceded by a AATAAA polyadenylation sequence (underlined in the figure). The amino acid sequence of the single long ORF is shown beneath the DNA sequence, beginning at the first in-frame ATG codon, which has a good match with the consensus sequence for translation start sites in *Drosophila* (Cavener, 1987). Exon boundaries were determined by comparison of cDNA and genomic sequences and are indicated above the sequence by the presence of numbers identifying the exons. Each of the six 36 amino acid repeats in the carboxyterminal part of the protein is indicated by a line beneath the amino acid sequence. This sequence has been deposited in the EMBL data library under accession number X62589.

insoluble aggregates (Gaul et al., 1987) from the appropriate *E. coli* cells and used as antigens.

Antisera directed against the pum protein were initially raised using pum polypeptide A. The specificities of antisera obtained from two rats were determined by western blot analysis using *Drosophila* ovarian proteins. Both sera detect a pair of proteins in the expected relative molecular mass range, but also react strongly with a protein that comigrates in SDS-polyacrylamide gels with an abundant yolk protein, as well as several other proteins (Fig. 7, lanes A and B). Such crossreactivity precludes use of the antisera to monitor position of pum protein in embryos, which contain high levels of yolk proteins. Given the extensive blocks of repeated amino acids in pum protein, it seemed possible that these regions might contribute to the crossreactivity. Consequently, to reduce or eliminate this problem, pum polypeptide B (which lacks many of the repeated blocks) was prepared and used for immunization. In western blot analysis using ovarian proteins each of two independent antisera preparations reacts specifically with the two proteins in the relative molecular mass range predicted for pum (Fig. 7, lanes C and D).

For western blot analysis, dissected ovaries were lysed in

SDS sample buffer, heated at 100° for 5 minutes, and centrifuged in a microfuge to pellet insoluble material. The supernatents were aliquoted into several tubes and frozen at −70°. After thawing and reheating at 100°, samples were electrophoresed in SDS-polyacrylamide gels. The gels were equilibrated in transfer buffer (48 mM Tris, 39 mM glycine, 20% methanol, 0.0375% SDS (pH 9.2)) for 15 minutes, and then transferred to nitrocellulose in a semi-dry transfer apparatus (Biorad). Blots were blocked and probed as described (Harlow and Lane, 1988), using alkaline phosphatase-coupled secondary antibodies from Jackson Immunoresearch Laboratories.

Immunohistochemical staining of fixed ovaries and embryos was performed as described (Macdonald and Struhl, 1986), using horseradish peroxidase-coupled secondary antibodies from Jackson Immunoresearch Laboratories. Only antisera specific for the pum protein, as measured by Western blot analysis, were used for the immunohistochemical analysis. Specificity of the primary antisera could not be confirmed by using a protein-null allele, since null alleles of the *pum* gene are zygotic-lethal (Lehmann and Nüsslein-Volhard, 1987).

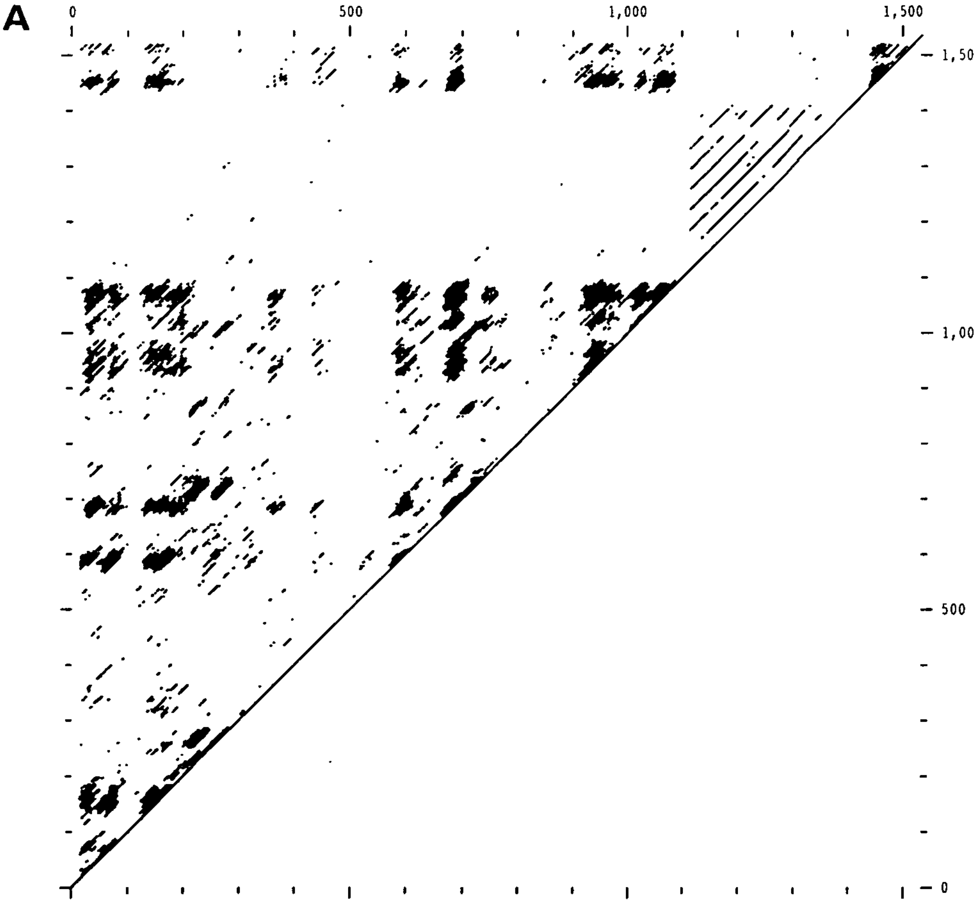
RNA analysis

In situ hybridizations were performed as described previously (Kim-Ha et al., 1991), using the whole-mount method of Tautz and Pfeifle (1989). The probe used for this analysis was a *Bgl*II-*Mlu*I restriction fragment (nucleotides 3598-4433 in Fig. 3).

Results

Genomic and cDNA clones of the pum *gene*

The *pum* locus has been mapped to a region of the third chromosome defined by the distal breakpoints in the overlapping deficiencies $Df(3R)p^{XT103}$ and $Df(3R)p^{XT9}$, which are pum^+ and pum^-, respectively (Lehmann and Nüsslein-Volhard, 1987). An inversion chromosome, *In(3R)Msc* (Lindsley and Grell, 1968), which breaks distally at about 85C and proximally within the cloned and well-characterized *Antennapedia* (*Antp*) complex (Scott et al., 1983; Kuroiwa et al., 1985), provides a convenient means to initiate a



B

```
                   *    *    *    *    *    *    *
     1111  DLANHIVEFSQDQHGSRFIQQKLERATAAEKQMVFS
     1147  EILAAAYSLMTDVFGNYVIQKFFEFGTPEQKNTLGM
     1183  QVKGHVLQLALQMYGCRVIQKALESISPEQQQEIVH
     1219  ELDGHVLKCVKDQNGNHVVQKCIECVDPVALQFIIN
     1255  AFKGQVYSLSTHPYGCRVIQRILEHCTAEQTTPILD
     1291  ELHEHTEQLIQDQYGNYVIQHVLEHGKQEDKSILIN
consensus  el gHv  L  DqyGnrVIQk LE  tpEq q i
```

Fig. 4. Repeat units in the *pum* gene. (A) Dot plot homology comparison of the predicted pum protein sequence aligned with itself. This comparison was generated using the COMPARE/DOTPLOT programs (Devereux *et al.*, 1984) with a window of 30 and a stringency of 15. (B) Alignment of the six copies of the 36 amino acid repeat unit. A consensus sequence is shown below the six repeats, with uppercase letters used to indicate residues present in 5 or 6 of the repeats, and lowercase letters to indicate residues present in 3 or 4 of the repeats. Note also that additional positions are highly conserved when conservative amino acid changes are considered.

chromosome walk in this region. Moreover, the *In(3R)Msc* chromosome is also *pum*$^-$ (Lehmann, 1985), and so the breakpoint falls within the *pum* gene. A genomic library was prepared from DNA of the inversion chromosome, and screened with a probe adjacent to the *Antp* breakpoint. Recombinant phage containing the inversion junction were identified and used to screen a wild-type genomic library to obtain unrearranged DNA from the 85C region. Subsequently, cosmid and phage libraries were screened to extend the walk to about 180 kb (Fig. 1), covering most or all of a *pum* transcription unit as described below.

To aid in the identification of the *pum* gene, several chromosomal positions defining its location were mapped. Distal breakpoints of *Df(3R)* p^{XT9} and *Df(3R)p*^{XT103} were mapped onto the walk by Southern blotting (Fig. 1; data not shown). The former deficiency, which is *pum*$^-$, removes a portion of the chromosome walk and must remove at least part of the *pum* gene. In contrast, the latter deficiency, which is *pum*$^+$, retains all of the region shown in Fig. 1, breaking at a position about 100 kb more proximal (Kim-Ha et al., 1991). In addition, the breakpoint of the inversion chromosome *In(3R)Msc* was mapped during the course of the cloning. This mutation simply inverts a large portion of the chromosome, and does not include any substantial DNA deletion in the 85C region (data not shown). Thus the inversion is apparently *pum*$^-$ because the breakpoint lies within the *pum* gene. In sum, these data indicate that the *pum* gene covers a region of at least 60 kb, extending into the region deleted in *Df(3R)* p^{XT9}, and crossing the *In(3R)Msc* breakpoint.

To identify potential *pum* transcripts, restriction fragments flanking the *In(3R)Msc* inversion breakpoint were labeled and used to probe northern blots of ovarian and early embryo RNA samples. Each of the initial fragments used detects the same pair of transcripts which appear in the expected temporal pattern: an abundant mRNA of about 6.5-7 kb, as well as a larger and rarer transcript (Fig. 2). Using the same probes, several cDNA clones were recovered from an ovarian cDNA library. Preliminary analysis of these cDNAs revealed that the mRNA was derived from an unusually long primary transcript, extending over a

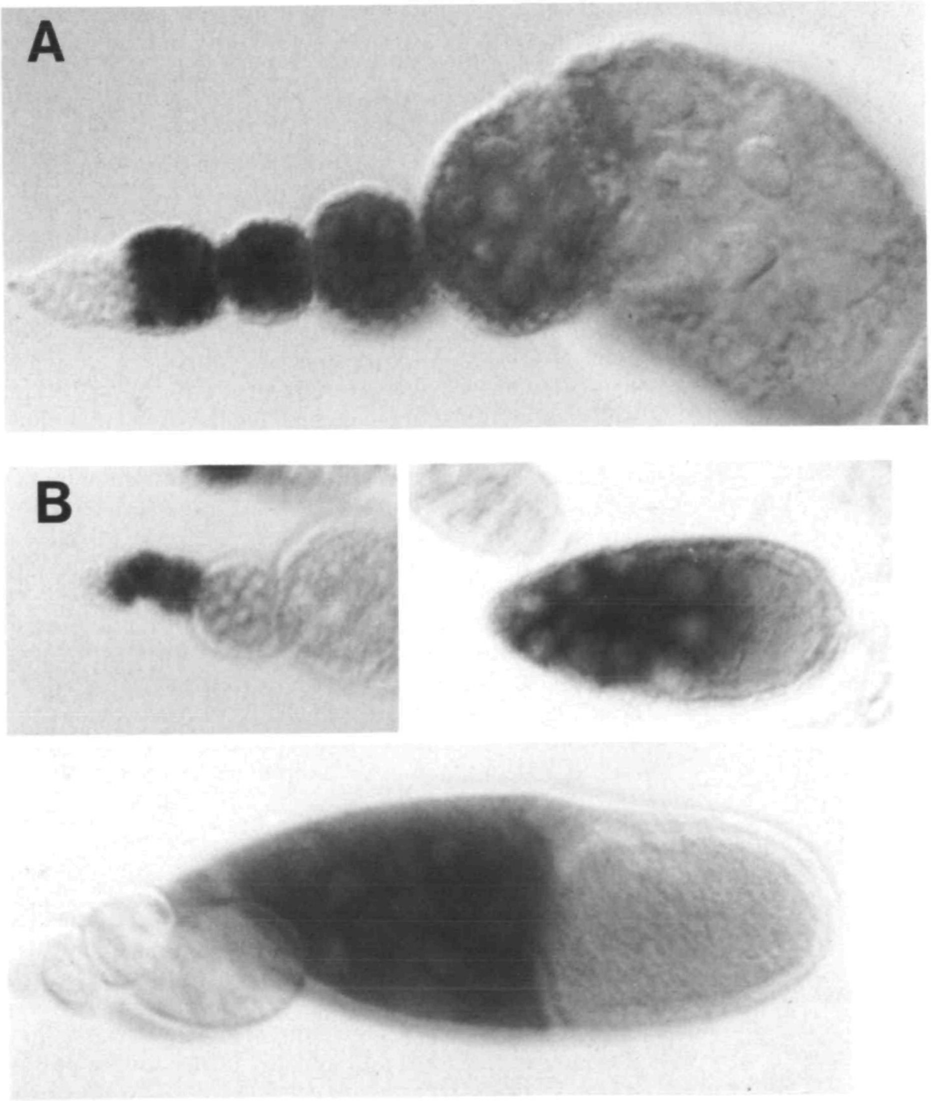


Fig. 5. Expression of *pum* mRNA during oogenesis. A and B show whole-mount preparations of ovarian tissue hybridized with a digoxigenin-labeled *Bgl*II-*Mlu*I fragment of the *pum* cDNA (from exons 10 and 11). The hybridization signal has been detected by an immunohistochemical method using alkaline phosphatase, and appears as darkly staining material on a light background. B is a photographic montage that includes the germarium and the early stages of oogenesis in the upper left part, a stage 9 egg chamber in the upper right part, and a stage 10 egg chamber in the lower part. *pum* transcripts are clearly present in the germarium and during early stages (1-4 or 5) of oogenesis. This is shown best in A, which is enlarged relative to B. *pum* mRNA is not detected during the intermediate stages 5 or 6-8 (A and B, upper left), but reappears by stage 9 (B, upper right) where it is present in the nurse cells. This pattern persists during stage 10 (B, lower).

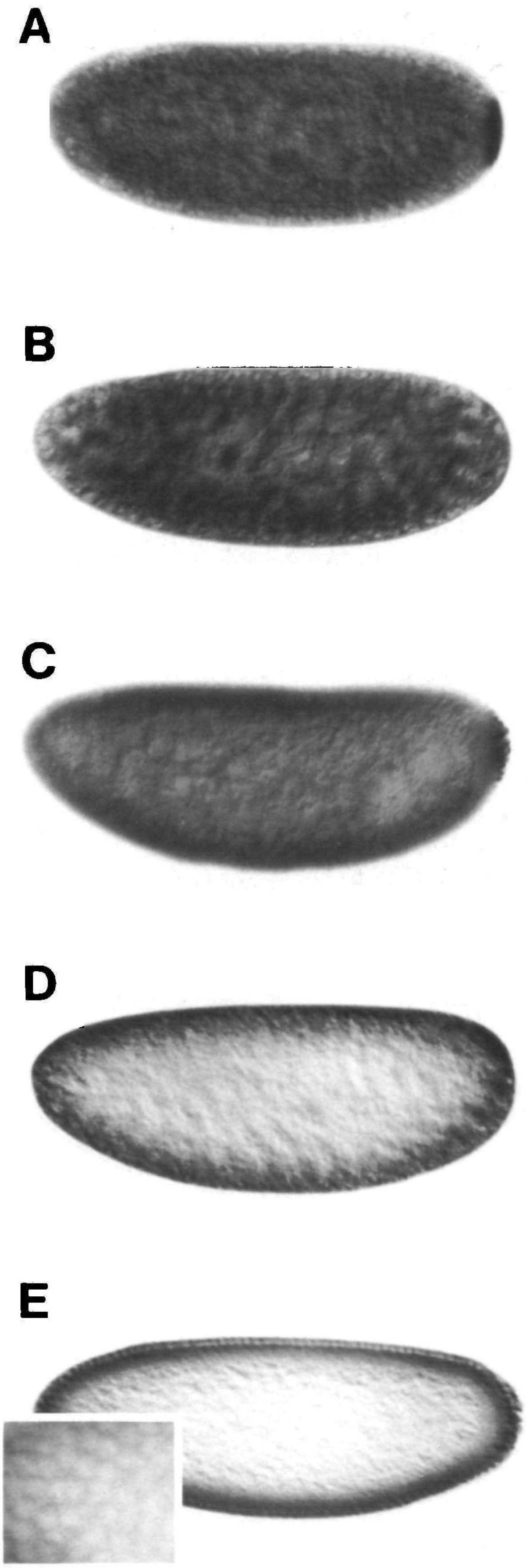


Fig. 6. Distribution of *pum* mRNA and protein in early embryos. Fixed embryos were processed to dectect *pum* mRNA (A-C) or pum protein (D-E). All embryos are oriented with posterior to the right and dorsal uppermost. A and B show *pum* mRNA in embryos prior to nuclear division cycle 10, when pole cells are formed. The embryo in C is from a later stage after pole cell formation. There is a high level of *pum* mRNA throughout each embryo. The embryos in A and C also display localized *pum* mRNA at the posterior pole. In C, the *pum* mRNA can be seen to be associated with the pole cells. D and E show embryos in which pum protein has been detected immunohistochemically. In the precellular blastoderm stage embryo in D, most of the staining is seen to be close to the cortex, with no asymmetry along the anteroposterior axis. At a later stage (E) the protein is now more closely associated with the cortical region, and is clearly seen to be cytoplasmic (inset - a surface view of the same embryo at greater magnification. The round nuclei show no staining while the cytoplasm is stained). Again, the staining pattern shows no asymmetry along the anteroposterior axis.

region of genomic DNA greater than 160 kb. Two types of evidence indicate that this transcript is derived from the *pum* gene. First, it fits the rather strict definition of *pum* gene location provided by the analysis of chromosomal rearrangements, namely, that it must span a region of over 60 kb, crossing the breakpoint of a *pum*$^-$ inversion chromosome, and extending into a region deleted in a *pum*$^-$ deficiency chromosome. The genomic inversion in *In(3R)Msc* interrupts the transcription unit [and the open reading frame (ORF)] between exons 5 and 6, and a large part of the transcript (again including part of the ORF) is deleted by *Df(3R)p*^{XT9}. In addition, the transcript remains intact in the *pum*$^+$ deficiency chromosome *Df(3R)p*^{XT103}. The second type of evidence is negative, in that no other 85C-region transcripts altered by the inversion have been detected using a variety of probes from the region.

Another criterion useful for supporting this identification of the *pum* transcript is its expected temporal pattern of expression. Mutants of *pum* display a maternal-effect phenotype (Lehmann and Nüsslein-Volhard, 1987), indicating that the gene must be expressed in ovaries. In addition there is a *pum*$^-$ adult phenotype (Lehmann and Nüsslein-Volhard, 1987) and so the gene is expected to be expressed during a later developmental stage as well. Northern blot analysis of the transcript indicates that it is expressed in ovaries and at a later stage of development, as expected (Fig. 2).

pum *gene structure*

Structure of the *pum* gene was determined in two steps. First, portions of the cDNAs were labeled and used as probes in Southern blot hybridization to cloned genomic DNA. This allowed an initial alignment of cDNA and genomic DNA sequences, and provided a rough approximation of the intron/exon organization. Second, the cDNA and much of the corresponding genomic DNA were sequenced (Fig. 3). None of the

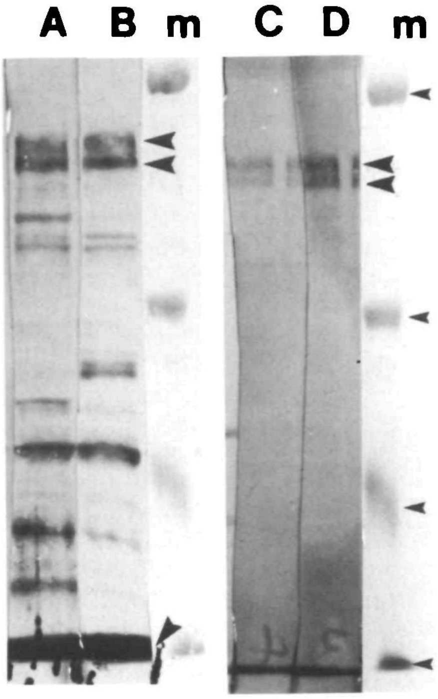


Fig. 7. Western blot detection of pum protein. Total ovarian proteins were separated by electrophoresis through a 5% polyacrylamide gel (30:0.8 acrylamide:bisacrylamide) in the presence of SDS, transferred to nitrocellulose filters, and probed with sera from rats immunized with pum polypeptide A (lanes A and B) or pum polypeptide B (lanes C and D). Bound antibodies were detected by enzyme activity staining after incubation with alkaline phosphatase-coupled secondary antibodies. The sera from both rats immunized with polypeptide A contain antibodies that bind to two protein bands that migrate at the position expected for pum protein, but also bind to several additional ovarian proteins. In contrast, the sera from both rats immunized with polypeptide B are specific for the two protein bands that migrate as expected for pum protein. It is not known if they represent two forms of pum protein. Protein relative molecular mass was estimated by comparison to prestained markers (Biorad). The size markers shown are 205, 116.5, 77, and 46.5 $\times 10^3$.

cDNAs were full length, and so parts of three independent clones were used to create a composite cDNA for sequencing. Although the composite cDNA does not extend to the 5′ end of the transcript (see Fig. 3 legend), it does include all of the single long open reading frame (ORF), since a stop codon is present in the cDNA upstream from the ORF in the same reading frame. It is not clear which of the two mRNAs the cDNAs correspond to, or if they are derived from the same species of transcript. Nevertheless, one of the individual cDNAs isolated includes all of the ORF, and so the predicted protein is a genuine product of the *pum* gene.

Unusual structure of the pum *protein*

Translation of the long ORF in the *pum* cDNA predicts a protein of 1533 amino acids with a predicted relative molecular mass of 157,413 (Fig. 3). There are two notable features of the protein, both of which are highlighted by a homology comparison of the protein sequence with itself (Fig. 4A). First, near the carboxy terminus there are six copies of a 36 amino acid repeat unit. Within the repeat unit 3 amino acids are invariant and 21 others are conserved (Fig. 4B). Second, much of the remainder of the protein consists of short regions in which a single amino acid is either highly enriched or present as a repeat. There are multiple domains of glycine, alanine and glutamine, and one domain highly enriched in serine and threonine at the carboxy terminus. Similar regions are found in many other *Drosophila* proteins, but rarely are the regions so extensive (Wharton et al., 1985; Haynes et al., 1987; Smoller et al., 1990).

Comparison of the protein sequence with the Genbank and EMBL data bases using the TFASTA program (Devereux et al., 1984) revealed no significant similarities. A further search was performed using the PROFILE program (Devereux et al., 1984), in which a weighted consensus sequence of the six copy repeat unit was used for comparison. Again, no significant similarities were detected.

pum *expression*

Because of the requirement for *pum* in the repositioning of the *nos* activity from the posterior pole of the embryo to its site of action in the presumptive abdominal region (Lehmann and Nüsslein-Volhard, 1987), it was of interest to determine if *pum* is expressed in a spatial pattern that might betray the nature of its role in that process. As an initial step in this analysis, the pattern of *pum* mRNA expression was visualized using a whole-mount in situ hybridization method (Tautz and Pfeifle, 1989). *pum* mRNA first appears in the most mature portions of the germarium (Fig. 5A, B). It is in the germarium that a cystoblast divides four times to yield 15 nurse cells and the oocyte, collectively referred to as an egg chamber. These divisions are incomplete, leaving the cells within a single egg chamber connected by ring canals. After the egg chambers separate from the germarium, they progress through a series of developmental stages defined by King (1970) and Mahowald and Kambysellis (1978). During the earliest stages, *pum* mRNA is present throughout the egg chamber, but subsequently disappears (Fig. 5A, B). Then, during stage 9, *pum* mRNA is expressed again, but the expression is now restricted to the nurse cells with little or no detectable mRNA in the oocyte (Fig. 5B). During stage 11, the contents of the nurse cells are transferred into the oocyte, and *pum* mRNA is presumably included in this movement. This transfer has not been visualized, since the whole-mount in situ hybridization method cannot be used for the stages following deposition of the vitelline membrane. Nevertheless, *pum* mRNA is present in early stage embryos (Fig. 6A,B), so the transfer does occur. The pattern of *pum* mRNA in early embryos is unusual in that it is not consistent among all embryos at a similar stage. Two patterns are observed among precellular blastoderm stage embryos: in some cases *pum* mRNA is spread throughout (Fig. 6B); in others *pum* mRNA is spread throughout, but there is also a posterior zone in which *pum* mRNA is further concentrated (Fig. 6A). Examples of each type of pattern are found in all of the intermediate stages prior to cellularization of the blastoderm. When *pum* mRNA is concentrated at the posterior pole of later stage embryos, it becomes associated with the pole cells (Fig. 6C). At present there is no good explanation for the variability in the pattern. However, it does not appear to be an experimental artefact. No variability was observed in parallel in situ hybridizations using probes specific for two tightly localized mRNAs, the anterior-specific *bicoid* transcript (Berleth et al., 1988) and the posterior-specific *oskar* transcript (Kim-Ha et al., 1991).

The issue of possible *pum* localization can also be addressed by visualizing the pattern of pum protein distribution. Antisera directed against bacterially produced pum protein fragments were prepared (see Materials and Methods) and used for immunohistochemical analysis of pum protein distribution in early stage embryos. There is no apparent asymmetry in the distribution of pum protein along the anteroposterior axis of the embryo at preblastoderm (Fig. 6D) or after the pole cells have formed (Fig. 6E), and no embryos are ever observed in which pum protein is more concentrated at the posterior pole. pum protein is present in cytoplasmic regions and not in the nuclei (Fig. 6E, inset), and is concentrated in the cortical region of the embryos beneath the nuclei (Fig. 6D,E).

Discussion

Organization of the posterior body pattern in *Drosophila* embryos requires the localized activity of the *nos* deteminant. A group of eight genes is necessary for posterior localization and activation of *nos*. One gene within this group, *pum*, is required during embryogenesis when *nos* activity apparently moves anteriorly from the posterior pole. As an initial step in exploring the role of *pum* in this process, the *pum* gene has been cloned, the sequence of its encoded product has been deduced, and the pattern of *pum* expression during the early stages of development has been determined.

Although the pum protein is not a homolog of any previously characterized protein, its amino acid sequence does have certain notable features. One is the prevalence of regions in which one of a number of single amino acids is highly enriched. Many *Drosophila* proteins contain homopolymeric runs of amino acids, such as *opa* sequences (Wharton et al., 1985) or *pen* repeats (Haynes et al., 1987), but these runs only rarely constitute such a large fraction of the protein (Smoller et al., 1990). For other *Drosophila* proteins, it has been proposed that these amino acid repeats are relatively 'unstructured', and may function as hinge or spacer regions separating different functional domains (Beachy et al., 1985; Laughon et al., 1985). If this is true for *pum*, then many of the functional domains may be small, or the spacer regions may be quite extensive. Some indication of the importance of these regions could come from phylogenetic comparisons of the *pum* gene. Unfortunately, the large size of the gene makes these experiments difficult to complete.

A second notable feature of the pum protein is the presence of six tandem copies of a 36 amino acid repeat unit. This repeat unit is easily detectable in a homology comparison, and includes three invariant amino acids as well as 21 other amino acids that are present in at least three of the repeat units. The presence of the repeats suggests a functional role but, since no similar repeat has been identified in any other protein, it is not yet possible to predict what their function might be.

Patterns of *pum* mRNA and protein accumulation have been examined to obtain further information about the mechanism of *pum* action. Surprisingly, variability is observed in mRNA distribution among individuals in a population of similarly staged embryos. In all cases, *pum* mRNA is distributed throughout the embryo, but in some embryos *pum* mRNA is also concentrated at the posterior pole. This difference could be due to variability in fixation, with loss of localization in some embryos. This seems unlikely since the embryos are fixed as a group and not individually. Another possibility is simply that the *pum* mRNA is not consistently localized to the posterior pole. Alternatively, posterior localization may be very dynamic and/or transient. Whatever the correct explanation may be, pum protein does not appear to be concentrated at the posterior pole, and is not localized asymmetrically along the anteroposterior axis. Since pum protein presumably provides the *pum* function, whatever *pum* mRNA localization does occur may not be important. Although pum protein is not localized along the body axis, it is restricted to the cytoplasm. Furthermore, pum protein is concentrated first in a cortical region in preblastoderm embryos, and later in a cortical region beneath the nuclei at the blastoderm stage. This region is therefore likely to be the site where *pum* influences *nos* movement or activity.

Given that we now know the amino acid sequence of the protein encoded by the *pum* gene and the distribution of the pum protein, what do we learn about the nature of the step for which *pum* is required, and the mechanism by which *pum* performs this function? Lehmann and Nüsslein-Volhard (1987) have suggested that *pum* is involved in transport of an abdominal determination signal, *nos*. Transport could occur along a cytoskeletal framework by the action of a molecular motor protein (Vale and Goldstein, 1990). The amino acid sequence of *pum* is not similar to any of the characterized molecular motor proteins, and thus is not likely to have that type of function. Movement of *nos* by such a mechanism could nevertheless be involved, and *pum* might be required to assemble or link *nos* to the transportation machinery. Alternatively, there may be no directed system for movement of *nos*. Rather, *nos* could move by diffusion (Crick, 1970), perhaps through an environment not conducive to its assembly, or promoting its inactivation or destabilization. In this model, *pum* might act to stabilize or protect *nos*, and could function in a manner similar to that proposed for chaperonins (Pelham, 1986). While the sequence of *pum* provides no direct support for this idea, the presence of extensive 'unstructured' regions in the pum protein could be involved in chaperonin-like activity. For any model of *pum* function, it is likely that some interaction with *nos* is involved, either direct or indirect. Thus the recent isolation of the *nos* gene (Wang and Lehmann, 1991) may facilitate biochemical studies that could lead to a more complete understanding of *pum* function

I thank Christianne Nüsslein-Volhard, Ruth Lehmann, Jim Kennison, and Kathy Mathews and the Indiana University stock center for fly stocks, Ron Kopito for identifying the 36 amino acid repeat in the *pum* coding region, Laura Kalfayan for providing her ovarian cDNA library, John Tamkun and Matt Scott for their cosmid library, Lex Bunten for assistance in preparing the manuscript, Gary Struhl and Robin Wharton for discussions, and the reviewers for their comments. This work was initiated while I was in Gary Struhl's lab at Columbia University and was supported there by the Howard Hughes Medical Institute. I particularly wish to thank Gary Struhl for giving me the opportunity to pursue this work in his lab. This work was supported by a grant from the NIH (GM42612). P.M.M. is a Pew Scholar in the Biomedical Sciences.

References

Beachy, P. A., Helfand, S. L. and Hogness, D. S. (1985). Segmental distribution of bithorax complex proteins during *Drosophila* development. *Nature* **313**, 545-551.

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. and Nüsslein-Volhard, C. (1988). The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J.* **7**, 1749-1756.

Boswell, R. E. and Mahowald, A. P. (1985). *tudor*, a gene required for assembly of the germ plasm in Drosophila melanogaster. *Cell* **43**, 97-104.

Cavener, D. (1987). Comparison of the consensus sequence flanking translation start sites in *Drosophila* and vertebrates. *Nucleic Acids. Res.* **15**, 1353-1361.

Crick, F. (1970). Difffusion in embryogenesis. *Nature* **225**, 420-422.

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

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**, 387-395.

Driever, W. and Nüsslein-Volhard, C. (1988a). The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell* **54**, 95-104.

Driever, W. and Nüsslein-Volhard, C. (1988b). A gradient of *bicoid* protein in Drosophila embryos. *Cell* **54**, 83-93.

Driever, W. and Nüsslein-Volhard, C. (1989). The bicoid protein is a positive regulator of *hunchback* transcription in the early *Drosophila* embryo. *Nature* **337**, 138-143.

Driever, W., Thoma, G. and Nüsslein-Volhard, C. (1989). Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the bicoid morphogen. *Nature* **340**, 363-367.

Frigerio, D., Burri, M., Bopp, D., Baumgartner, S. and Noll, M. (1986). Structure of the segmentation gene *paired* and the Drosophila PRD gene set as a part of a gene network. *Cell* **47**, 735-746.

Frohnhöfer, H. G., Lehmann, R. and Nüsslein-Volhard, C. (1986). Manipulating the anteroposterior pattern of the *Drosophila* embryo. *J. Embryol. exp. Morph.* **97** *Supplement* 169-179.

Frohnhöfer, H. G. and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the *Drosophila* embryo by the maternal gene *bicoid*. *Nature* **324**, 120-125.

Gaul, U., Seifert, E., Schuh, R. and Jäckle, H. (1987). Analysis of *Kruppel* protein distribution during early Drosophila development reveals posttranscriptional regulation. *Cell* **50**, 639-647.

Harlow, E. and Lane, D. (1988). *Antibodies A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Haynes, S. R., Rebbert, M. L., Mozer, B. A., Forquignon, F. and Dawid, I. B. (1987). *pen* repeat sequences are GGN clusters and encode a glycine-rich domain in a *Drosophila* cDNA homologous to the rat helix destabilizing protein. *Proc. natn. Acad. Sci. U.S.A.* **84**, 1819-1823.

Kim-Ha, J., Smith, J. L. and Macdonald, P. M. (1991). *oskar* mRNA is localized to the posterior pole of the Drosophila oocyte. *Cell* **66**, 23-35.

King, R. C. (1970). *Ovarian Development in* Drosophila melanogaster. New York: Academic Press.

Kuroiwa, A., Kloter, U., Baumgartner, P. and Gehring, W. J. (1985). Cloning of the homeotic *Sex combs reduced* gene in *Drosophila* and *in situ* localization of its transcripts. *EMBO J.* **4**, 3757-3764.

Laughon, A., Carroll, S. B., Storfer, F. A., Riley, P. D. and Scott, M. P. (1985). Common properties of proteins encoded by the Antennapedia complex genes of *Drosophila melanogaster. Cold Spring Harbor Symp. Quant. Biol.* **50**, 253-262.

Lehmann, R. (1985). Regionsspezifische Segmentierungsmutanten bei *Drosophila melanogaster* Meigen. Tübingen.

Lehmann, R. and Nüsslein-Volhard, C. (1986). Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of *oskar*, a maternal gene in Drosophila. *Cell* **47**, 141-152.

Lehmann, R. and Nüsslein-Volhard, C. (1987). *hunchback*, A gene required for segmentation of an anterior and posterior region of the *Drosophila* embryo. *Devl Biol.* **119**, 402-417.

Lindsley, D. L. and Grell, E. H. (1968). Genetic variations of *Drosophila melanogaster. Carnegie Inst. Wash. Publ.* 627.

Macdonald, P. M. and Struhl, G. (1986). A molecular gradient in early *Drosophila* embryos and its role in specifying the body pattern. *Nature* **324**, 537-545.

Mahowald, A. P. and Kambysellis, M. P. (1978). Oogenesis. In *Genetics and Biology of Drosophila, Vol. 2*, (ed. M. Ashburner and T.R.F. Wright), pp. 141-224. New York: Academic Press.

Manseau, L. and Schüpbach, T. (1989). *cappuccino* and *spire*: two unique maternal-effect loci required for both the anterioposterior and dorsoventral patterns of the *Drosophila* embryo. *Genes Dev.* **3**, 1437-1452.

Nüsslein-Volhard, C., Frohnhöfer, H. G. and Lehmann, R. (1987). Determination of anteroposterior polarity in *Drosophila. Science* **238**, 1675-1681.

Pelham, H. R. B. (1986). Speculations on the functions of the major heat shock and glucose-regulated proteins. *Cell* **46**, 959-961.

Rosenberg, A. H., Lade, B. N., Chui, D., Lin, S., Dunn, J. J. and Studier, F. W. (1987). Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene* **56**, 125-135.

Sander, K. and Lehmann, R. (1988). *Drosophila* nurse cells produce a posterior signal required for embryonic segmentation and polarity. *Nature* **335**, 68-70.

Scott, M. P., Weiner, A. M., Hazelrigg, T. I., Polisky, B. A., Pirrotta, V., Scalenghe, F. and Kaufman, T. C. (1983). The molecular organization of the *Antennapedia* locus of Drosophila. *Cell* **35**, 763-776.

Smoller, D., Friedel, C., Schmid, A., Bettler, D., Lam, L. and Yedvobnick, B. (1990). The *Drosophila* neurogenic locus *mastermind* encodes a nuclear protein unusually rich in amino acid homopolymers. *Genes Dev.* **4**, 1688-1700.

Struhl, G., Struhl, K. and Macdonald, P. M. (1989). The gradient morphogen *bicoid* is a concentration-dependent transcriptional activator. *Cell* **57**, 1259-1273.

Tautz, D. and Pfeifle, C. (1989). A non radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals a translational control of the segementation gene hunchback. *Chromosoma* **98**, 81-85.

Vale, R. D. and Goldstein, L. S. B. (1990). One motor, many tails: An expanding repertoire of force-generating enzymes. *Cell* **60**, 883-885.

Wang, C. and Lehmann, R. (1991). Nanos is the localized posterior determinant in Drosophila. *Cell* **66**, 637-647.

Wharton, K. A., Yedvobnick, B., Finnerty, V. G. and Artavanis-Tsakonas, S. (1985). Opa: A novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in *D. melanogaster*. *Cell* **40**, 55-62.

Zinn, K., DiMaio, D. and Maniatis, T. (1983). Identification of two distinct regulatory regions adjacent to the human beta-interferon gene. *Cell* **34**, 865-879.

(*Accepted 21 September 1991*)