The EMBO Journal vol.6 no.3 pp.791-799, 1987

Structure and sequence of the *Drosophila zeste* gene

V.Pirrotta[1], E.Manet[2,3], E.Hardon[1,4], S.E.Bickel[1] and M.Benson[1]

[1]Department of Cell Biology, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA, and [2]EMBL, Postfach 102209, Heidelberg, FRG

[3]Present address: Laboratoire d'epidemiologie et immunovirologie des tumeurs Faculte de Medecine Alexis Carrel, Rue G.Paradin, 69372 Lyon, France

[4]Present address: EMBL, Postfach 102209, Heidelberg, FRG

Communicated by V.Pirrotta

The *zeste* gene of *Drosophila* affects the expression of other genes in a manner that depends on the homologous pairing of the chromosomes bearing the target gene. *Zeste* mediates transvection effects, the ability of one gene to control the expression of its homologous copy on another chromosome. We have determined the structure of the *zeste* gene and several mutants bearing partial deletions and the sequence of the z^+, z^1, z^{op6} and z^{11G3} alleles. The predicted *zeste* protein has an unusual structure including runs of Gln, Ala and alternating Gln Ala. Contrary to expectations the z^1, z^{op6} and z^{11G3} mutations can each be attributed to single amino acid changes. The analysis of the mutants suggests that the *zeste* gene product is required for normal expression of at least some genes and we argue that z^a mutants may have residual function.

Key words: transfection/chromosome pairing/gene regulation

Introduction

The *zeste* gene of *Drosophila* is a regulatory gene that affects the expression of certain other loci. Three of these have been identified: *white*, *Ultrabithorax* (*Ubx*) and *decapentaplegic* (*dpp*) although a systematic search for interactions with other target loci has not been undertaken.

Genetic tests for zeste interaction are based on two phenomena. One is transvection (Lewis, 1954; Kaufman *et al.*, 1973) a phenomenon whereby certain alleles of a given locus, in heterozygous combination are able to complement one another partially, provided that they are juxtaposed by somatic chromosome pairing and that an active *zeste* gene is present. *Zeste*-dependent transvection effects hve been demonstrated at *Ubx*, *dpp* and *white* (Babu and Bhat, 1980; Gelbart and Wu, 1982) but pairing-dependent allelic complementation may well be a more widespread phenomenon. As the molecular structure of the genes involved has become better understood, it is becoming clear that the combinations of alleles that transvect are those involving one regulatory mutation and one mutation affecting the transcribed part of the target gene. It appears, in other words, that transvection reveals the ability of an intact regulatory region on one copy of the target gene to control the expression of the other copy, whose regulatory region is mutationally altered. This interpretation rationalizes the requirement for the physical proximity of the two copies that result from homologous pairing of the chromosomes. It does not clarify the involvement of the *zeste* gene product which is not itself required for chromosome pairing.

The second genetic test for *zeste* function is based on the properties of the original *zeste* mutation, isolated by Gans (1953) and referred to as z^1 in this paper. The z^1 mutant product affects specifically the function of the *white* gene, greatly depressing its expression when *white* is present in two paired copies. This effect occurs at the level of transcription (Bingham and Zachar, 1985) and is not appreciably detected if a single copy of *white* is present. The target of the z^1 product has been localized to a 95-nucleotide interval in the regulatory region of *white*, about 1100 bp distant from the transcription start site. This target can be brought closer or moved away from the *white* promoter by at least 800 bp or inverted with respect to it without altering its effect (Pirrotta *et al.*, 1985). The z^1 mutation is antagonized by the z^+ product, providing therefore a simple genetic test for z^+ function. Mutations of *zeste* that fail to complement z^1 are generally called z^a-type. However, z^1 is partially dominant over z^+ and this dominance is enhanced and revealed by the presence of increasing numbers of copies of *white*, either paired or tandemly repeated (Jack and Judd, 1979). Evidence that the *zeste* product also interacts with a single copy of the target gene is provided by the z^{op6} mutant, isolated by Lifschytz and Green (1984) starting from the z^1 mutant. The z^{op6} and other mutants of that class now give the *zeste* effect even with a single copy of the *white* gene. This effect is enhanced, and z^{op6} becomes dominant over z^+, when the *white* gene is paired. The vast majority of *zeste* mutations are of the z^a variety and behave as hypomorphic or null alleles with respect to both z^1 complementation and transvection.

The two aspects of *zeste* function revealed by transvection and by the z^1 effect on *white* are clearly related in some way. Both involve the *zeste* gene, occur at the RNA level and depend on pairing but there is not always a correspondence between the effects of *zeste* mutations on transvection and on *white* expression. For example, the z^1 allele behaves like z^+ in promoting transvection at *Ubx*; z^{11G3}, a pseudo-revertant of z^1 is nearly wild-type in complementing z^1 effects at *white* but behaves like z^- with respect to transvection at *Ubx* (Kaufman *et al.*, 1973). Worse yet, it appears that different genes respond differently: z^1 permits transvection at *Ubx* but not at *dpp* (Gelbart and Wu, 1982) and finally, even the *white* gene is affected by z^1 in a tissue-specific manner: its expression is depressed in the eye but not in the ocelli, testes or malpighian tubules (Gans, 1953; Pirrotta *et al.*, 1985; Bingham and Zachar, 1985). This plethora of different effects suggests that *zeste* acts in processes that involve multiple components and may well differ in mechanistic details at different sites.

It is important to emphasize that both transvection and the z^1 effect on *white* involve artificial situations that reveal the participation of *zeste* but do not tell us what, if anything, *zeste* normally does to the expression of genes. To answer this question we need more direct information on the *zeste* product and its molecular interactions with other nuclear components. The *zeste* gene has been cloned (Mariani *et al.*, 1985; Gunaratne *et al.*, 1986). In the present work we have analysed the structure of the

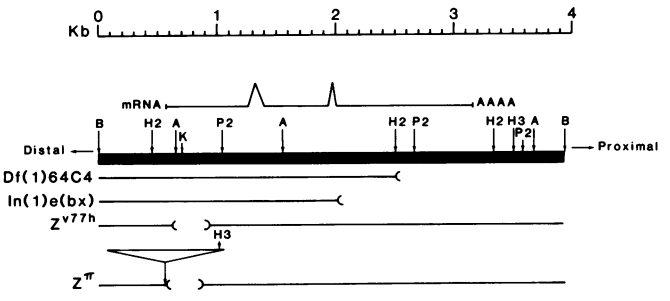

Fig. 1. Map of the *zeste* gene. The figure summarizes the results of different mapping experiments. The top line shows the scale in kilobases. The second line represents the transcript. The 5′ end is on the left and the position of introns is indicated. The heavy black bar represents the DNA with the relevant restriction sites symbolized by: B, *Bam*HI; H2, *Hin*dII; A, *Ava*I; K, *Kpn*I; P2, *Pvn*2; H3, *Hin*dIII. The lines below show the breakpoints of *Df(1) 64c4* and *In(1) e(bx)* and the deletions and insertion present in z^{v77h} and z^{π}. Proximal and distal refer to the orientation in the X chromosome.

gene, its transcription and its sequence. We have compared the wild-type with a number of the most significant *zeste* mutants as a prelude to the study of the *zeste* protein and its function.

Results

Germ line transformation

We have shown in a previous article that several *zeste* mutations [*In(1) e(bx)*], z^{π}, z^{v77h}, etc.] are associated with breakpoints or alterations in a 3.9-kb *Bam*HI fragment whose map is shown in Figure 1 (Mariani *et al.*, 1985). To determine whether this fragment contains the entire *zeste* gene, we decided to use the corresponding region from the z^{op6} mutant. If the fragment contains a functional gene, it should confer a yellow eye phenotype to z^- flies. Furthermore, both males and females should be yellow-eyed since the *white* gene is on the X chromosome and since the distinctive property of the z^{op6} product is that it affects even a single, unpaired copy of *white*. We inserted the z^{op6} *Bam*HI 3.9-kb fragment in the pUChsneo vector for P-mediated germ line transformation and injected the DNA into $y\ z^a$ embryos, carrying a defective *zeste* gene, unable to complement the z^1 mutation. All the transformants detected in the F1 generation by the criterion of G418 resistance had a pale yellow eye color, irrespective of sex, indicating that they had received a functional z^{op6} gene. Five independent transformed lines behaved identically, with no evidence of position effects.

To begin mapping the lesions in the mutant gene, we constructed a hybrid *zeste* transposon. We took the 5′ end of the gene, from the *Bam*HI site to the *Kpn*I site, from z^{op6} while the rest of the gene, from the *Kpn*I site on, came from the wild-type. The G418-resistant F1 flies obtained from $y\ z^a$ embryos all had red eyes and gave rise to no yellow-eyed flies upon inbreeding. Since the *Kpn*I site is well within the transcribed region (see below), this experiment tells us that the z^1 component of z^{op6} lies not in the 5′ flanking region but in the transcribed sequences. We can draw no conclusions about the second component of z^{op6} which allows it to affect a single copy of *white*, since we cannot detect it independently of z^1.

Sequence of the zeste region

We have sequenced most of the *Bam*HI 3.9-kb fragment as well as the homologous fragments isolated from three important mutant alleles that alter the function of the *zeste* product: the z^1 mutant of Gans, its pseudo-revertant z^{11G3} (Gans, 1953) and the z^{op6} mutant of Lifschytz and Green (1984). The z^1 mutant was isolated from a wild-type stock that can be expected to contain several sequence polymorphisms with respect to the Oregon R sequence that represents our wild-type. Both the z^{11G3} revertant and the z^{op6} mutant were isolated from z^1 flies and should share with the z^1 sequence all those polymorphisms including the z^1 mutation proper and contain in addition new sequence changes responsible for the z^{11G3} or z^{op6} phenotype. The sequence obtained is shown in Figure 2 which also summarizes the data from cDNA sequencing and S1 mapping of the transcripts. We will defer for the moment a discussion of these results.

A number of other mutants alter the structure of this region and affect the function of the *zeste* gene. One of these is *Df 64c4*, a deficiency that spans the interval from *zeste* to the 3C region and fails to complement the z^1 mutation. Genomic S1 mapping experiments such as those shown in Figure 3 show that this deficiency has a breakpoint near position 2500 ± 10 in the map shown in Figure 1. Genomic Southern blots show that *In(1) e(bx)* has a breakpoint in the segment represented by probe A1 (Figure 3b) and S1 mapping shows that the breakpoint is near position 2015 ± 5. This mutation, an inversion with a breakpoint in the *zeste* gene, behaves as a z^- allele both by failing to complement z^1 and by failing to promote transvection effects (Lewis, 1954; Kaufman *et al.*, 1973).

The two other mutants studied were both induced by P−M dysgenic crosses. One, z^{π}, contains an insertion of P element sequences, while the other, z^{v77h}, has no P sequences associated with it but has instead a small deletion of ~300 nucleotides (Green, 1984; Mariani *et al.*, 1985). Genomic S1 mapping experiments place the z^{v77h} deletion approximately between nucleotides 629 ± 5 and 933 ± 5 and show that the z^{π} mutant bears a deletion of similar size between position 570 ± 4 and 865 ± 5, in addition to the P element insertion.

In the case of the z^{π} mutation, a genomic clone containing the *zeste* gene and the inserted P element was recovered by microdissection from mutant chromosomes (Mariani *et al.*, 1985) and the precise position of the insertion and the extent of the flanking deletion were determined by nucleotide sequencing. The results, summarized in Figures 1 and 2, confirm the interpretation of the genomic S1 mapping experiments and give the following picture. In z^{π} a P element ~1 kb long is inserted at position 568. Restriction mapping data and DNA sequence agree that this is the 3′ end of the P element. The 5′ end of the P element is contiguous with the *zeste* sequence at position 867. The deletion of 299 nucleotide pairs and the absence of the target site duplication that commonly accompanies transposon integration make it impossible to determine whether the insertion originally occurred at position 568 or at position 867 or even somewhere in between and was then followed by a small bilateral deletion.

The z^{v77h} mutation was almost surely caused by a P element insertion at the same site. In this case the insertion was followed by a deletion that removed the entire P element along with some 300 nucleotides of flanking sequence. Neither breakpoint coincides with the breakpoints of the z^{π} mutation and again it is impossible to determine whether the deletion was to the right or to the left of the original insertion or bilateral.

S1 mapping of zeste mRNA

We utilized a series of overlapping fragments from the *zeste* region to map, by S1 protection, the sequences represented in *zeste* RNA from adult flies. Representative experiments are shown in Figure 4. It is clear that, under the conditions used in most of the experiments, some of the duplex regions were not entirely immune to S1 nuclease, resulting in a number of minor bands. The results, however, indicate the presence of three exons:

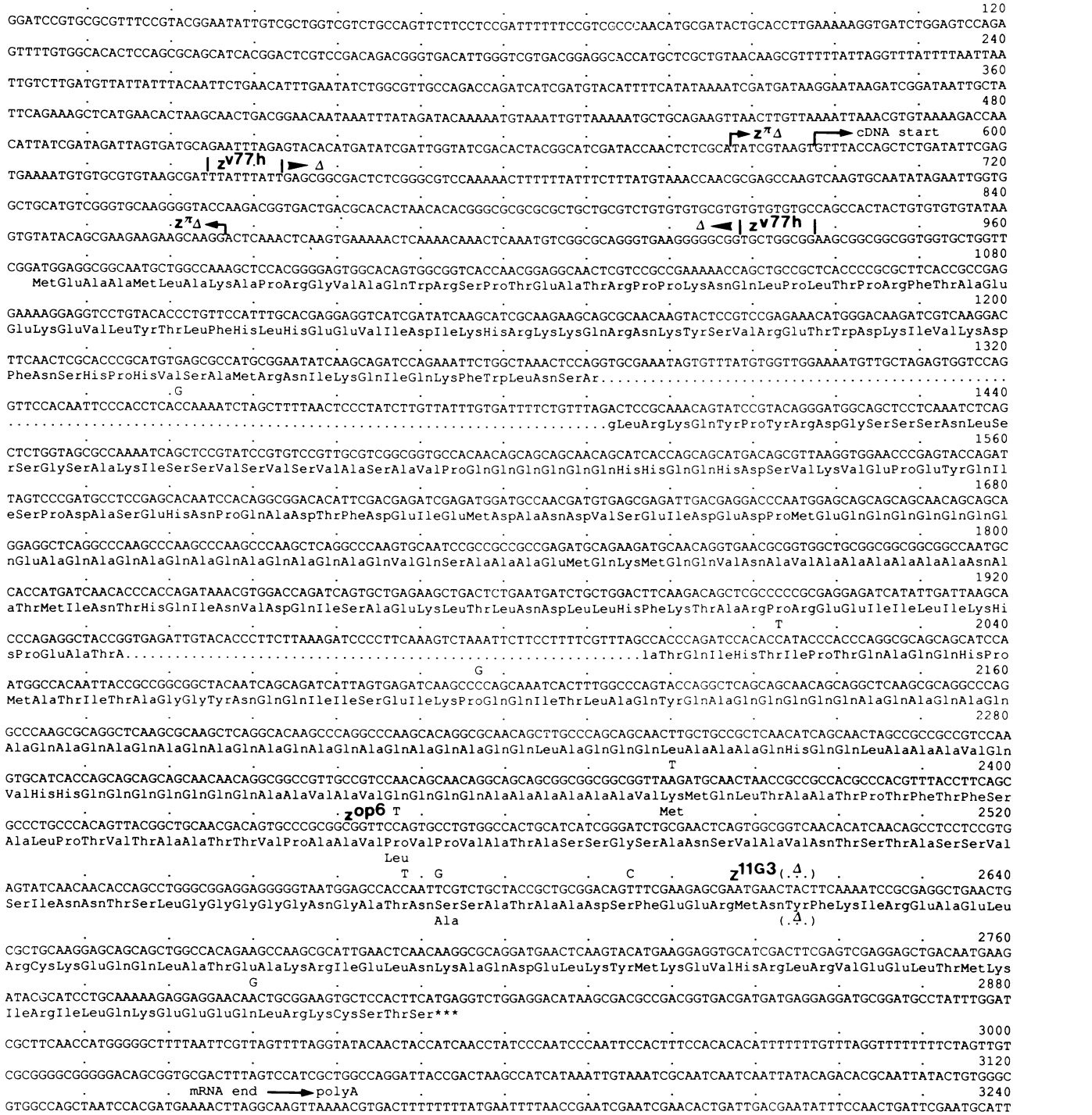

Fig. 2. Sequence of the *zeste* gene. The nucleotide sequence starts at the *Bam*HI site at position 1 on the left in Figure 1. The position of the insertion–deletion in z^{π} was determined by sequencing the z^{π} clone. The deletion breakpoints in z^{v77h} were determined by genomic S1 mapping and are shown with uncertainty limits. The nucleotide changes found in the z^{1} sequence are shown above the sequence and the corresponding amino acid changes, if any, below the amino acid sequence.Some of these changes are shared by a cDNA clone isolated from wild-type. The z^{op6} and z^{11G3} sequences contain all of the changes found in z^{1} plus the changes specifically attributed to these two mutants.

one from position 572 ± 4 to position 1270 ± 2, corresponds to the largest protected fragment in Figure 3 (panel labelled H1). A second exon begins at position 1392 ± 2 and ends at position 1935 ± 4. A third exon begins at position 2000 ± 4 and ends at position 3160 ± 5. The first and the third exons are more sensitive to internal attack in the S1 protection experiments, giving

rise to partial products that can be mapped with some confidence. When the sensitive regions are identified in the nucleotide sequence in Figure 1, it is evident that in most cases, they correspond to stretches of high A+T content flanked by sequences of high G+C content. Examples of these S1 sensitive regions are those at positions 603, 623, 656, 883, 891 and 902 that give rise to the series of bands visible in Figure 3, Panel H1, and those at positions 3169, 2987, 2973 and 2901 that are responsible for minor bands in Figure 3, Panel A1. The intensity of these minor bands was variable in different experiments, indicating that the internal cleavage is dependent on reaction conditions.

S1 mapping was also carried out with various mutant RNAs. No difference in signal intensity and in exon size or placement was detectable with z^1, z^{op6} or z^a RNA (not shown) indicating that the mutations do not affect overall transcript abundance or splicing pattern. S1 mapping of *In(1) e(bx)* RNA shows that it is normal up to the second intron but lacks detectable third exon sequences, in agreement with the genomic mapping data.

Northern blot analysis shows that both the z^{π} and the z^{v77h} mutants make an RNA that is ~300 nucleotides shorter than normal (Mariani *et al.*, 1985). S1 mapping of the RNA from these mutants shows that the deletion affects the first exon. For z^{v77h}, both the RNA analysis and the genomic DNA analysis agree in placing the deletion fully within the first exon which simply becomes shorter. For z^{π} the P insertion site coincides or even precedes by a few nucleotides the position of the 5′ end of the wild-type RNA. In this mutant, *zeste* exon sequences are first detected beginning with position 867, in precise agreement with the end of the deletion flanking the P element. This implies that in z^{π}, the *zeste* RNA includes some P element sequences. Since the RNA is shorter than normal by 200−300 nucleotides, this result also implies that either the normal transcription start site is used but most of the P element sequence is spliced out, or that transcription begins with the P element but includes only a very small amount of P element sequence. In either case the orientation of this transcript is opposite to that of the normal endogenous P element transcript. Interestingly, upon dysgenic crosses, the z^{π} mutation can give rise to partial revertants as well as mutants with more severe phenotypes (manuscript in preparation).

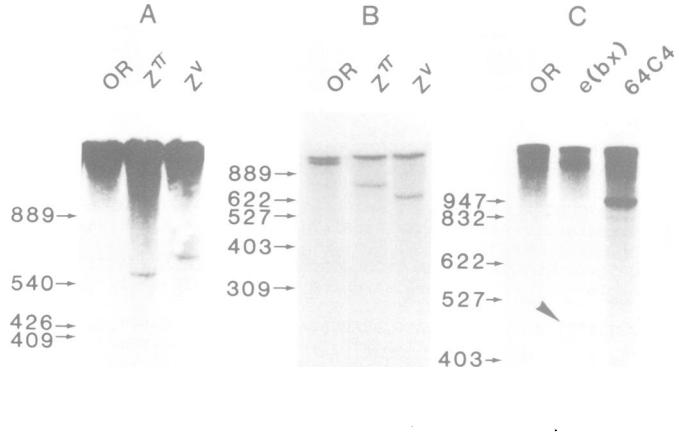

Fig. 3. Genomic S1 mapping of *zeste* mutants. Genomic DNA from different *zeste* mutants was denatured and allowed to hybridize to a single-stranded probe from different regions of the *zeste* gene, digested with S1 nuclease and the protected fragments hybridized on denaturing acrylamide gels. In these experiments and most of those shown in Figure 4, residual undigested probe is generally represented by the top band. **(A)** Genomic DNA from Oregon R (OR), z^{π} or z^{v77h} flies protected by a probe corresponding to the entire *Bam*HI 3.9-kb fragment. **(B)** The same three genomic DNAs protected by probe A2 in Figure 4b. **(C)** Genomic DNA from Oregon R, *In(1) e(bx)* or *Df 64c4* flies protected by probe A1 in Figure 4b. The small fragment, indicated by the arrowhead, shows the existence of a breakpoint in *e(bx)* near one end of the probed fragment. A similar experiment using the *Bam*HI 3.9-kb fragment as probe located the position of the breakpoint unambiguously (not shown).

cDNA clones

When we screened cDNA libraries to isolate *zeste* cDNA clones, we found that as many as 1% of the clones in an adult cDNA library gave significant hybridization to non-overlapping probes from the *zeste* region. Upon further analysis, it became clear that these signals were due to homology with the central part of the *zeste* sequence, which contains a very high degree of reiteration of the trinucleotides CAG and CAA. When sequenced, representatives of these cDNA clones in fact contained long stretches of repeated CAG triplets and no other sequence homologous to *zeste*.

More realistic results were obtained using probes from the 3′ or the 5′ end of the gene. With these probes we isolated five cDNA clones and sequenced their extremities. In all cases the 3′ end, after subtracting the poly(A) tail, corresponded to the sequence ending at position 3157, in excellent agreement with the S1 mapping results. The longest clone was just over 2400 nucleotides long, corresponding to a full length cDNA sequence. The 3′ sequence of this clone ended with 12 A residues preceded by the genomic sequence ending at position 3157. The 5′ end

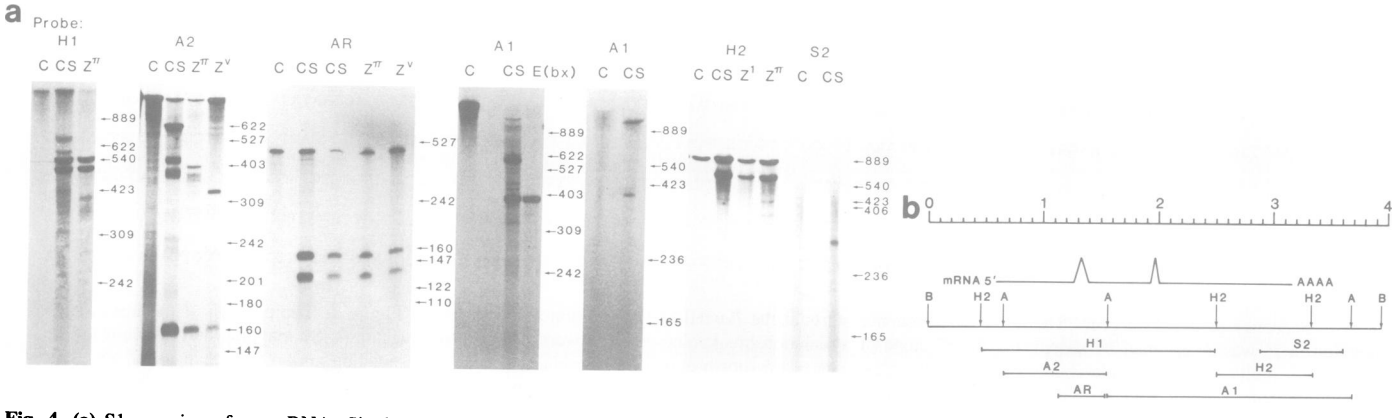

Fig. 4. (a) S1 mapping of *zeste* RNA. Single-stranded labelled probes made from m13 clones were hybridized to poly(A)$^+$ RNA from different mutants digested with S1 nuclease and analyzed on 4% acrylamide−urea gels. The extents of probes H1, A2, AR, A1, H2 and S2 are shown in Figure 4b. In each experiment, a control (C) is shown, in which the probe was mock-hybridized to tRNA. The two panels labelled A1 show that under different S1 digestion conditions it is possible to obtain protection of the entire third exon. C, tRNA control; CS, wild-type Canton S RNA; z^{π}, z^{π} RNA; z^{v}, z^{v77h} RNA; E(bx), *In e(bx)* RNA. The position of mol. wt size markers is indicated. **(b)** S1 mapping probes.

began with 12 T residues, followed by 24 nucleotides that are the inverted complement of the 24 nucleotides preceding the poly(A) tail at the 3′ end. The genuine 5′ sequence of the cDNA begins with the GTTT at position 578 in the genomic sequence. This corresponds, within a few nucleotides, to the 5′ end predicted by the S1 protection experiments. It is likely, therefore, that the cap site for *zeste* mRNA corresponds to one of the nucleotides immediately preceding position 578 and that the inverted repeat at the 5′ end was produced by an artefact during the construction of the cDNA library.

A plausible account for the production of the inverted repeat envisages the annealing of the 3′ end of the first cDNA strand to its own 5′ end (corresponding to position 3130–3138). The 3′ end was then extended to copy the sequence at the 5′ end (the 3′ end of the RNA), including the 12 Ts corresponding to the poly(A) tail. The resulting 12 As then served as a priming site for oligo(dT) to synthesize a complete second strand.

Whether or not this scenario explains the artifact, the resulting cDNA clone confirms the placement of the 3′ and 5′ ends of the RNA. We determined the sequence of most of this cDNA clone and found that the position of the intron–exon boundaries is strictly consistent with the S1 mapping data, with the restriction map of the cDNA clone and with the presence of consensus splicing donor and acceptor sites at the expected positions.

The length of the mRNA deduced from these results is 2398 nucleotides, exclusive of the poly(A) tail. This is an excellent agreement with our estimate of 2.4–2.5 kb from Northern blot hybridization results. We cannot exclude the possible existence of an additional 5′ exon provided it is so small (<50 nucleotides) as to have gone undetected in our S1 mapping experiments.

Discussion

Transcription start and polyadenylation site

S1 mapping results and the sequence of the cDNA clone suggest that *zeste* transcription starts at or near position 578 but the region immediately preceding this site contains no recognizable TATA sequence. A perfect TATAAA is found at position 321 but there are no plausible splicing donor or acceptor sequences consistent with the existence of a short 5′ exon in the region between 340 and 578. Not all *Drosophila* genes have recognizable TATA sequences. *White* and *Ubx* are two genes that do not have one. In the case of *zeste*, as in that of *white*, the low levels of mRNA observed might lead us to expect a non-standard promoter sequence.

If transcription starts near position 578, the sequence AGTGTT surrounding this site would be in excellent agreement with the cap sequence of several other *Drosophila* genes shown in Table I and resembles the consensus proposed by Hultmark *et al.* (1986): ATCATG/TTT/C. We note also that although no TATA sequence precedes the presumptive cap site, the immediate upstream region contains repeated sequence motifs of the form TATCGATA, sometimes arranged in a larger inverted repeat structure such as that centered between positions 544 and 545. These might have a regulatory significance.

At the 3′ end of the gene, the processing/polyadenylation site is established by the sequence of five cDNA clones at position 3157 where four As in the genomic sequence merge with the poly(A) tail. It is noteworthy that here also *zeste* diverges from the norm in lacking the AATAAA almost universally found 15–30 nucleotides before the polyadenylation site (Proudfoot and Brownlee, 1976). No sequence withn 60 nucleotides of this site could be converted to AATAAA by a single base change. The closest approach to this sequence is provided by the AATCCA at 3132 or the GATGAA at 3139.

The predicted zeste protein

The first 400 nucleotides of the transcribed region contain six ATGs. The first three are followed by terminators within a few codons. The fourth runs into a terminator at the beginning of the second exon. The fifth and sixth ATGs are in the same reading frame and only four codons apart. We cannot determine at present which of these two is selected as the initiator codon but if we take the first, at position 964, we obtain an open reading frame that, after splicing out the two intron sequences, runs until position 2811 and codes for a polypeptide of 555 amino acids and 61 102 mol. wt. Codon usage in this reading frame is in excellent agreement with *Drosophila* practice. The length of the untranslated leader and the presence of several AUG codons before the actual translation start suggest the possibility of translational regulation. Certainly this region as well as the coding sequence are rich in clusters of very high G+C content with potential for strong secondary structure. Notably, a large part of this leader sequence is deleted in the z^{π} and z^{v77h} mutants.

The predicted *zeste* protein (Figure 5), although not unusual in its overall amino acid composition, has an uncommon distribution of residues. The amino acid sequence falls readily into four sections. The amino terminal section is rich in both basic and acidic residues with positive charges predominating. A second region, starting with residue 154 is strongly acidic. The third region, consisting of the third exon up to residue 481, is almost devoid of basic or acidic residues, is slightly hydrophobic overall and contains a large number of Gln and Ala. Finally the carboxy terminal region is again densely packed with basic and acidic residues. These regions coincide roughly with domains suggested by secondary structure calculations. Figure 6 shows that the second and fourth regions are expeced to be predominantly helical, separated by regions of predominantly random coil conformation. The protein is predicted to have very little β-sheet content.

The most striking feature of the amino acid sequence is the presence of extensive runs of Gln, of Ala and of alternating Gln and Ala. Frequently these runs form a boundary between regions of distinct secondary structure or amino acid composition. A particularly long stretch of Gln and Ala, mostly in alternation (position 326–403) forms an extended bridge predicted to be helical between two regions of random coil. These runs correspond to those regions of the sequence that cross-hybridize to many

Table I. Comparison of selected *Drosophila* RNA cap sequences

putative *zeste*	AGTGTTT
yellow	AGTCGTT
per	AGTGTTC
yp-1	AGTTCAA
yp-2	ATGCAGTACAA
ddc	AGTTAAG
α-amylase	CAGAGTGAAA
ftz	AGGGCTC
sgs-8	GGTTACC
Antp I	TTGATAGGAGTCGTA
Antp II	TTCAGTTGTG
EIP 28	CATCAGTTCAG

The first nucleotide on the left is the reported 5′ end. Sequences are aligned to emphasize the homology. Sources are: *yellow*: Geyer *et al.* (1986); *per*: Jackson *et al.* (1986); *yp-1* and *yp-2*: Hung and Wensink (1983); *ddc*: Eveleth *et al.* (1986); α-amylase: Boer and Hickey (1986); *ftz*: Laughon and Scott (1984); sgs-8: Garfinkel *et al.* (1983); *Antp*: Schneuwly *et al.* (1986); EIP 28: Cherbas *et al.* (1986).

```
z-protein
                          10                  20                  30                  40
MetGluAlaAlaMetLeuAlaLysAlaProArgGlyValAlaGlnTrpArgSerProThrGluAlaThrArgProProLysAsnGlnLeuProLeuThrProArgPheThrAlaGluGlu

                          50                  60                  70                  80
LysGluValLeuTyrThrLeuPheHisLeuHisGluGluValIleAspIleLysHisArgLysLysGlnArgAsnLysTyrSerValArgGluThrTrpAspLysIleValLysAspPhe
                                                  Intron I
                          90                 100                 110                 120
AsnSerHisProHisValSerAlaMetArgAsnIleLysGlnIleGlnLysPheTrpLeuAsnSerArgLeuArgLysGlnTyrProTyrArgAspGlySerSerSerAsnLeuSerSer

                         130                 140                 150                 160
GlySerAlaLysIleSerSerValSerValSerValAlaSerAlaValProGlnGlnGlnGlnGlnGlnHisHisGlnGlnHisAspSerValLysValGluProGluTyrGlnIleSer

                         170                 180                 190                 200
ProAspAlaSerGluHisAsnProGlnAlaAspThrPheAspGluIleGluMetAspAlaAsnAspValSerGluIleAspGluAspProMetGluGlnGlnGlnGlnGlnGlnGlnGlu

                         210                 220                 230                 240
AlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnValGlnSerAlaAlaAlaGluMetGlnLysMetGlnGlnValAsnAlaValAlaAlaAlaAlaAlaAlaAsnAlaThr

                         250                 260                 270                 280
MetIleAsnThrHisGlnIleAsnValAspGlnIleSerAlaGluLysLeuThrLeuAsnAspLeuLeuHisPheLysThrAlaArgProArgGluGluIleIleLeuIleLysHisPro
     Intron II
                         290                 300                 310                 320
GluAlaThrAlaThrGlnIleHisThrIleProThrGlnAlaGlnGlnHisProMetAlaThrIleThrAlaGlyGlyTyrAsnGlnGlnIleIleSerGluIleLysProGlnGlnIle

                         330                 340                 350                 360
ThrLeuAlaGlnTyrGlnAlaGlnGlnGlnGlnGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnAlaGlnGlnLeu

                         370                 380                 390                 400
AlaGlnGlnGlnLeuAlaAlaAlaGlnHisGlnGlnLeuAlaAlaAlaValGlnValHisHisGlnGlnGlnGlnGlnGlnAlaAlaValAlaValGlnGlnGlnGlnAlaAlaAla
                                                                                  op6
          Met            410                 420                 430      z   Leu            440
AlaAlaAlaValLysMetGlnLeuThrAlaAlaThrProThrPheThrPheSerAlaLeuProThrValThrAlaAlaThrThrValProAlaAlaValProValProValAlaThrAla

                         450                 460                 470                 Ala 480
SerSerGlySerAlaAsnSerValAlaValAsnThrSerThrAlaSerSerValSerIleAsnAsnThrSerLeuGlyGlyGlyGlyGlyAsnGlyAlaThrAsnSerSerAlaThrAla
                    11G3
                  z     (...)               500                 510                 520
AlaAspSerPheGluGluArgMetAsnTyrPheLysIleArgGluAlaGluLeuArgCysLysGluGlnGlnLeuAlaThrGluAlaLysArgIleGluLeuAsnLysAlaGlnAspGlu

                         530                 540                 550
LeuLysTyrMetLysGluValHisArgLeuArgValGluGluLeuThrMetLysIleArgIleLeuGlnLysGluGluGluGlnLeuArgLysCysSerThrSerEnd
```

Fig. 5. Amino acid sequence of the *zeste* protein. Amino acid changes for the z^1 mutant are shown above the wild-type sequence. The z^{op6} and z^{11G3} proteins include all the z^1 changes plus the changes specifically indicated. Glutamine and alanine residues in the repetitive regions are emphasized by stippling.

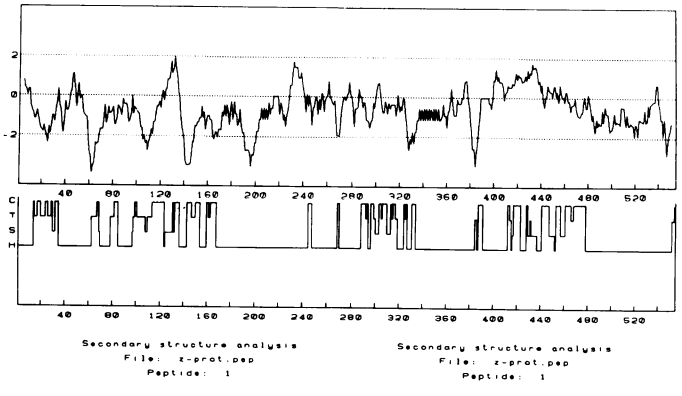

Fig. 6. Hydropathicity and secondary structure plots. The top diagram shows the hydropathicity profile calculated by the method of Kyte and Doolittle (1982) using a window of nine amino acids. Negative values correspond to hydrophilic regions. The lower diagram represents the secondary structure predictions made by the method of Garnier *et al.* (1978). C, coil; T, turn; S, β-sheet; H, α-helix. The ordinate indicates the residue number.

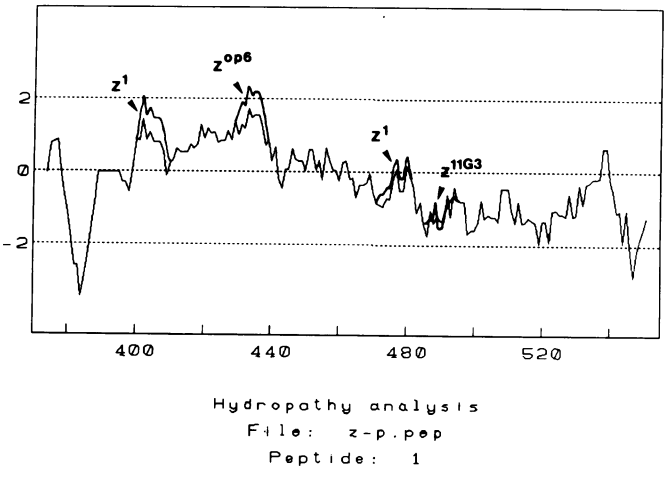

Fig. 7. Hydropathicity comparisons for *zeste* mutants. The profiles shown are calculated as in Figure 6 for the last 185 residues of the wild-type and mutant proteins. For z^{11G3} a shift of one unit in the ordinate was made at position 490 to allow for the deletion of one amino acid.

disperse sites in the genome (Mariani *et al.*, 1985) and to many and abundant sequences in the cDNA libraries. These sequences are in fact related to the *opa* sequence in the *Notch* gene (Wharton *et al.*, 1985), to the M repeat in the *Ubx* and *Antennapedia* genes (McGinnis *et al.*, 1984; Schneuwly *et al.*, 1986) and to similar stretches in the *engrailed* gene (Poole *et al.*, 1985). The significance of their frequent recurrence in recently sequenced, developmentally important genes is unknown. Perhaps they simply occur frequently in the *Drosophila* genome and it is simply a coincidence that most of the recently analyzed *Drosophila* genes are of developmental importance. It is remarkable, however, that runs of glutamines have been found recently in the amino acid sequence of several eukaryotic regulatory proteins such as the rat glucocorticoid receptor (Miesfeld *et al.*, 1986) or the human c-*myc* oncogene (Colby *et al.*, 1983). These glutamines may play a role in interactions with other nuclear components. We note for example the possibility that glutamine rich regions may be

subject to cross linkage to the ε amino group of lysine residues on the same or on different molecules by the action of transglutaminases (Falk, 1980).

We have screened the NBRF and PIR protein sequence data bases for homologies to *zeste*. When the Gln- and Ala-rich regions were subtracted, we could not detect significant homology to known proteins. The *zeste* protein does not possess structures obviously resembling the helix-turn-helix configuration characteristic of many DNA binding domains (Pabo and Sauer, 1984; Ohlendorf *et al.*, 1985). Amino acid loops (zinc fingers), held together by interactions with divalent cations, have been shown to be another structural principle common to another class of gene regulatory proteins (Miller *et al.*, 1985; Rosenberg *et al.*, 1986; Hartshorne *et al.*, 1986). *Zeste* does not contain cysteine and histidine residues in the spacing expected for such loops. Other proteins known to bind specifically to DNA lack such recognizable structural motifs. If *zeste* binds to DNA, it either manages to form equivalent configurations from less obvious constituents or it utilizes a different structural principle.

Zeste mutants

A comparison of the wild-type *zeste* sequence with that of z^1 shows, as expected, the presence of numerous polymorphisms, not surprisingly concentrated in the middle, repetitive part of the gene. No changes were found in the 5′ flanking region or in the untranslated leader sequence, consistent with the germline transformation results. Most of the changes are either in intron sequences or in the third position of codons and do not affect the predicted amino acid sequence. A C to G change at position 1341 in the first intron generates a CCTCAGCA that constitutes a possible splicing acceptor site. Splicing at this site would have drastic consequences for the z^1 protein since a terminator would be encountered after two codons. However, S1 mapping of z^1 RNA rules out this possibility since the second exon is identical in size to that of the wild-type and begins at the same site. Some of the polymorphisms, including a T to G change at position 2572 that replaces Ser 276 with an alanine, were found to be present also in the sequence of the full length cDNA clone. Since the cDNA library was constructed from an Oregon R population, we conclude that these changes, including the Ser to Ala change, are not significant. It is possible that silent changes have significant effects on the biological activity of the mRNA. It would be very difficult, however, to explain the z^1 phenotype in these terms since z^1 behaves as if it had an altered activity, antagonistic to that of z^+. There remains one important nucleotide change: an A to T at position 2360 that results in the replacement of Lys 405 with a methionine. This mutation, in the repetitive middle region of the protein, removes the only charged residue in a stretch of 160 amino acids and is likely to have a significant effect on the properties of the *zeste* product (Figure 6). We note that this lysine occurs in the tripeptide Lys-Met-Gln at the end of an extended Gln-Ala region. It may be significant that the same tripeptide occurs also earlier at position 224 in the amino acid sequence, following a shorter run of Gln and Ala. The z^{op6} mutant contains all the polymorphisms found in the z^1 sequence as well as the Lys to Met change, as expected, since z^{op6} was derived from z^1 by ethylmethane sulphonate mutagenesis (Lifschytz and Green, 1984). In addition, a C to T mutation at position 2447, changes Pro 234 into a leucine. This mutation also is located in the relatively hydrophobic middle region of the protein and, like the z^1 mutation, has the effect of increasing its hydrophobic character. The pseudorevertant z^{11G3} was derived from z^1 by X-irradiation (Gans, 1953). Its sequence is identical to that of z^1 except for a deletion of three nucleotides: the repeated triplet ACTACT at position 2612 is reduced to a single ACT. The consequence is the deletion of Tyr 490 from the amino acid sequence.

The simplest hypothesis for *zeste* function is that it acts on the DNA either by binding directly to regulatory sequences or by interacting with other proteins that bind to DNA. We now know that the *zeste* product is a DNA binding protein that recognizes specific sequences (M.Benson and V.Pirrotta, manuscript in preparation). The phenotypes of the z^1 and z^{op6} mutants indicate that their products are still able to interact with their target genes but are altered with respect to other interactions required for the expression of the *white* gene, z^{op6} more so than z^1. We suppose, therefore, that the central, relatively hydrophobic domain of the protein normally interacts with other proteins. These could be structural components of chromosomes or nucleus, other *trans*-acting factors, or other *zeste* monomers.

The genetic experiments using z^1/z^+ or z^{op6}/z^+ heterozygotes suggests that the proximity of additional copies of the *white* gene gives the z^1 or z^{op6} product an advantage over the z^+ protein in binding to their target. A plausible mechanism that would explain this advantage would be one in which the mutant protein is better able to form an oligomer which would now have two DNA binding sites, able to bind to two copies of the genes at the exclusion of other interactions that might be required of the normal *zeste* product.

Many interpretations of the z^{11G3} revertant are possible at present. This revertant acts nearly like the wild-type in antagonizing the z^1 effect at *white* but it is defective in promoting transvection at *Ubx*. One interesting possibility is that z^{11G3} might be less efficient in DNA binding. In this case, it would still be able to form mixed oligomers with the z^1 product but the oligomer would not be bifunctional and would lack the pairing dependent enhancement. Needless to say, these interpretations are at present purely speculative and are still inadequate to account for all the observations.

z^a-type mutants

The mutants we want to consider here are those apparently *zeste* null mutants caused by DNA rearrangements. These are the *Df 64c4, In(1) e(bx)*, z^π and z^{v77h}. All of these are unable to complement z^1 with respect to its effect on *white*.

We note that both z^π and z^{v77h} are affected only in the untranslated leader sequence and possess an otherwise normal protein coding region. We interpret these mutants as hypomorphs, defective in translational efficiency, possibly in translational controls, but otherwise able to produce normal *zeste* product. The presence of at least some *zeste* activity is shown by the fact that both z^{v77h} and z^π support detectable transvection effects at *Ubx* (Green, 1984; V.Pirrotta, unpublished observations). It is of interest to note that both mutants show abnormalities in eye pigmentation. These are more evident in z^{v77h} which has a diluted eye color that turns brownish with age in both males and females (hence independent of pairing) and is furthermore non-uniform or variegated. The effect in the z^π mutant is less pronounced but easily noticeable. These mutants, which can only be interpreted as underproducers of a normal *zeste* product, suggest strongly that the *zeste* product is necessary for normal *white* expression.

z^a-type mutants are generally said to have normal eye pigmentation. This assertion is however not borne out by inspection in all cases. The original z^a mutant (Gans, 1953) has a distinctly diluted eye color in both males and females that turns towards brown with age. The effect on eye pigmentation is not observed

in all z^a-type mutants. A conspicuous case is that of *In(1) e(bx)* which has eyes indistinguishable from wild-type. This mutant was originally isolated for its inability to support transvection at *Ubx* and was later shown not to complement z^1. The *e(bx)* mutant lacks about 800 nucleotides of the *zeste* coding region but contains an otherwise normal 5′ half of the gene. In particular we know that *In(1) e(bx)* makes an RNA that terminates shortly after the inversion breakpoint (Mariani *et al.*, 1985) hence does not add on to the protein extraneous amino acids. We propose that the *e(bx)* mutant protein has residual activity that is adequate for some of the functions of the normal *zeste* product, including normal *white* gene expression, but not adequate to compete with z^1 product or to permit the pairing-dependent interaction required for transvection. We note furthermore that the *64c4* deletion, which is also considered z^- because it does not complement z^1 contains a nearly full length *zeste* gene, lacking only the last 300 nucleotides of the coding sequence. This mutant too may have residual *zeste* activity. Finally we would like to raise again the possibility that *zeste* may be an essential gene. The genetic evidence points to a negative answer. Arguments such as the preceding ones, however, indicate that the classical z^a mutants are not necessarily null mutants. The genetic tests for *zeste* function are indirect and measure distinctly different properties of the *zeste* product. The genetic evidence itself and the results reported here suggest that *zeste* is involved in multiple interactions that differ in detail at different loci. Certain kinds of mutations may affect some functions of *zeste* without abolishing all. The answer to this question will require a direct molecular analysis of the *zeste* product and its function as well as the generation of true *zeste* null mutations in which the coding sequence of the gene is substantially deleted.

Materials and methods

Germ line transformation

The *Bam*HI 3.9-kb fragment isolated from the z^{op6} and z^+ fragments was cloned in the *Bam*HI site of the pUChsneo vector (Steller and Pirrotta, 1985). The cloned DNA was injected into *y* z^a embryos at a concentration of 400 μg/ml, together with 100 μg/ml helper plasmid phs$^\pi$ DNA (Steller and Pirrotta, 1986). Surviving adults were crossed to uninjected *y* z^a partners and the progeny selected on food containing G418. This was made by rehydrating 1.37 g of instant food, formula 4-24 (Carolina Biological Supply Co.) with 5 ml of water containing 800 μg/ml G418 (Geneticin, obtained from Sigma).

Construction and screening of mutant genomic libraries

To isolate the *zeste* gene from different mutants, genomic *Eco*RI libraries were constucted by ligating 1 μg of *Eco*RI-cut genomic DNA with 2–5 μg of *Eco*RI-cut λNM 1149 phage DNA (Murray, 1983). The ligated DNA was packaged as described by Scherer *et al.* (1981) and plated directly for screening without amplification. Filters (Benton and Davis, 1977) were hybridized using a small single-stranded probe corresponding to the non-repetitive 3′ end of the *zeste* gene. To isolate cDNA clones, we used a cDNA library made from adult flies generously provided by B.Yedvobnick and screened it with the same probe. The full length cDNA clone was selected by double-screening the positive clones with a second probe made from first exon sequences.

DNA sequencing

We used fragments cloned in the mp8, mp9 or mp18 vectors (Messing, 1983) and the dideoxy chain termination method of Sanger *et al.* (1977). The wild-type sequence was determined on both strands using overlapping restriction fragment subclones in both orientations as well as partial exonuclease deletion clones. The mutant sequences were determined on one strand only except for a few regions. Progressive deletions from one end were produced by exonuclease III digestion by the method of Henikoff (1984). The unusual wealth of G+C-rich clusters occasionally produced secondary structure artifacts such as band compression and anomalous migration in the sequencing gels in a few places, to resolve ambiguities, we resorted to synthetic oligonucleotide primers, and to the use of reverse transcriptase instead of the Klenow fragment of DNA polymerase. The sequence was assembled and later analysed on a Sperry personal computer connected through a network to the Baylor Molecular Biology Information Resource.

S1 mapping

Single-stranded probes were made from m13 subclones by priming the phage DNA and synthesis in the presence of radioactive nucleotide triphosphates. The probes were hybridized to RNA or genomic DNA and S1 digested as described by Pirrotta and Bröckl (1984). When the S1 digestion was done at 37°C, some regions of the RNA–DNA hybrid in the first and third exon were prone to internal attack. This could be reduced by digesting at 25°C but frequently at the cost of greater background.

Acknowledgements

We are grateful to Christa Garber for her skill in S1 mapping, to Elise Koster and Elaine McGuffin for technical assistance and to Darla DiStefano for typing the manuscript. We are indebted to Ting Wu and Mel Green for mutant strains, to Juan Codina for preparing the oligonucleotides, to Barry Yedvobnick for the cDNA library and to Charlie Lawrence and Dan Goldman for computer wizardry. E.M. was the recipient of an EMBO fellowship, E.H. was supported in part by an EMBL predoctoral fellowship and S.E.B. by an NSF predoctoral fellowship. The research was financed by a grant to V.P. from the US National Institutes of Health.

References

Babu,P. and Bhat,S.G. (1980) In Siddigi,O., Babu,P., Hall,L.M. and Hall,J.C. (eds), *Development and Neurobiology in Drosophila*. Plenum Press, New York, pp. 35–40.
Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–182.
Bingham,P.M. and Zachar,Z. (1985) *Cell*, **40**, 819–825.
Boer,P.H. and Hickey,D.A. (1986) *Nucleic Acids Res.*, **14**, 8399–8412.
Cherbas,L., Schulz,R.A., Koehler,M.M., Savakis,C. and Cherbas,P. (1986) *J.Mol. Biol.*, **189**, 617–631.
Colby,W., Chen,E., Smith,D. and Levinson,A. (1983) *Nature*, **301**, 722–725.
Eveleth,D.D., Gietz,R.D., Spencer,C.A., Nargany,F.E., Hodgetts,R.B. and Marsh,J.L. (1986) *EMBO J.*, **5**, 2663–2672.
Falk,J.E. (1980) *Annu. Rev. Biochem.*, **49**, 517–531.
Gans,M. (1953) *Bull. Biol. Fr. Belg.*, **38** (Suppl.), 1–90.
Garfinkel,M.D., Pruitt,R.E. and Meyerowitz,E.M. (1983) *J. Mol. Biol.*, **168**, 765–789.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.
Gelbert,W.M. and Wu,C.T. (1982) *Genetics*, **102**, 179–189.
Geyer,P.K., Spana,C. and Corces,V.G. (1986) *EMBO J.*, **5**, 2657–2662.
Green,M.M. (1984) *Mol. Gen. Genet.*, **194**, 275–278.
Gunaratne,P.H., Mansukhani,A., Lipari,S.E., Liou,H.-C., Martindale,D.W. and Goldberg,M.L. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 701–705.
Hartshorne,T.A., Blumberg,H. and Young,E.T. (1986) *Nature*, **320**, 283–287.
Henikoff,S. (1984) *Gene*, **28**, 351–359.
Hultmark,D., Klemenz,R. and Gehring,W.J. (1986) *Cell*, **44**, 429–458.
Hung,M.-C. and Wensink,P.C. (1983) *J. Mol. Biol.*, **164**, 481–492.
Jack,J.W. and Judd,B.H. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 1368–1372.
Jackson,F.R., Bargiello,T.A., Yun,S.-H. and Young,M.W. (1986) *Nature*, **320**, 185–188.
Kaufman,T.C., Tasaka,S.E. and Suzuki,D.T. (1973) *Genetics*, **75**, 299–321.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Laughon,A. and Scott,M.P. (1984) *Nature*, **310**, 25–31.
Lewis,E.B. (1954) *Am. Nat.*, **88**, 225–229.
Lifschytz,E. and Green,M.M. (1984) *EMBO J.*, **3**, 999–1002.
Mariani,C., Pirrotta,V. and Manet,E. (1985) *EMBO J.*, **4**, 2045–2052.
McGinnis,W., Levine,M.S., Hafen,E., Kuroiwa,A. and Gehring,W.J. (1984) *Nature*, **308**, 428–433.
Messing,J. (1983) *Methods Enzymol.*, **101**, 20–78.
Miesfeld,R., Rusconi,S., Godowski,P.J., Maler,B.A., Okret,S., Wikström,A.C., Gustafsson,J.-Å. and Yamamoto,K.R. (1986) *Cell*, **46**, 389–399.
Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **4**, 1609–1614.
Murray,N.E. (1983) In Hendrix,R.W., Roberts,J.W., Stahl,F.W. and Weinberg,R.A. (eds), *Lambda II*. Cold Spring Harbor Laboratory, NY, pp. 395–432.
Ohlendorf,D.H., Anderson,W.F. and Matthews,B.W. (1983) *J. Mol. Evol.*, **19**, 109–114.
Pabo,C.O. and Sauer,R.T. (1984) *Annu. Rev. Biochem.*, **53**, 293–321.
Pirrotta,V. and Bröckl,C. (1984) *EMBO J.*, **3**, 563–568.
Pirrotta,V., Steller,H. and Bozzetti,M.P. (1985) *EMBO J.*, **4**, 3501–3508.
Poole,S.J., Kauvar,L.M., Drees,B. and Kornberg,T. (1985) *Cell*, **40**, 37–43.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211–214.
Rosenberg,V.B., Schroder,C., Preiss,A., Kienlin,A., Cote,S., Riede,I. and Jackle,H. (1986) *Nature*, **319**, 336–339.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*,

74, 5463–5467.
Scherer,G., Telford,J., Baldari,C. and Pirrotta,V. (1981) *Dev. Biol.*, **86**, 438–447.
Schneuwly,S., Kuroiwa,A., Baumgartner,P. and Gehring,W. (1986) *EMBO J.*, **5**, 733–739.
Steller,H. and Pirrotta,V. (1985) *EMBO J.*, **4**, 167–171.
Steller,H. and Pirrotta,V. (1986) *Mol. Cell. Biol.*, **6**, 1640–1649.
Wharton,K.A., Yedvobnick,B., Finnerty,V.G. and Artavanis-Tsakonas,S. (1985) *Cell*, **40**, 55–62.

Received on November 12, 1986; revised on December 29, 1986