The retrotransposon *copia* controls the relative levels of its gene products post-transcriptionally by differential expression from its two major mRNAs

Cathy Brierley and Andrew J.Flavell*
Department of Biochemistry, The University of Dundee, Dundee DD1 4HN, UK

Received January 9, 1990; Revised and Accepted April 11, 1990

ABSTRACT

All retroviruses and retrotransposons studied to date regulate the relative levels of *gag* and *pol/int* gene products post-transcriptionally from a single mRNA. In these genetic elements the production of protein encoded by the *pol* and *int* genes is attenuated by a translational stop or frameshift in the reading frame preceding their coding regions in the mRNA. We show here that the *Drosophila* retrotransposon *copia* also produces lower amounts of gene products from its *int/pol* region than *gag* region but this is achieved by a mechanism which is novel for this class of genetic element. We show by the use of gene fusion constructs that the subgenomic 2 kilobase *copia* RNA, encoding *gag* products, is expressed as protein in cultured cells at least ten-fold more efficiently than the full genome length RNA, which additionally contains the *pol* and *int* open reading frames.

INTRODUCTION

Retroviruses and retrotransposons are members of a family of transposable genetic elements present in the genomes of all eukaryotes which have been searched thoroughly for them (1−3). These genetic elements all share a transposition mechanism which involves the transcription of the integrated genomic DNA copy into an RNA of nearly full length which contains all of the genetic information of the transposable element. This RNA is found in the virus, or in the virus-like particle of retrotransposons and is copied by reverse transcription into extrachromosomal DNA which becomes inserted into new chromosomal locations. The full length RNA is also the template for both the protein components of the virion core (collectively termed *gag* proteins) and the enzymes involved in the replication process, namely the protease, reverse transcriptase and integrase.

Retroviruses, as distinct from retrotransposons, specify at least one additional RNA which encodes the glycoproteins found in the viral membrane envelope. This latter RNA is a spliced subgenomic species with the same 5′ end as the full length RNA but containing its own translational initiation and termination signals (4). Other more complicated retroviruses, such as HIV 1, contain additional spliced subgenomic RNAs but while these are required for efficient replication, they are thought to encode proteins which are not directly involved in the transposition process (5−7). Neither these additional RNAs, nor the envelope glycoprotein-encoding RNA are thought to exist in the most basic retrovirus-related transposable elements, the retrotransposons.

Retroviral and retrotransposon proteins are needed in differing relative amounts for the efficient replication of the mobile elements encoding them. The virus (or virus-like) particle contains many more copies of the core structural components than of the enzymes catalysing precursor polyprotein cleavage, reverse transcription and integration (4). These proteins are all derived by proteolytic processing of a precursor polypeptide whose translation is initiated from a single site on the full length RNA (4). Low levels of *pol/int* gene products, relative to *gag* products, are produced from this single RNA by the presence of a translational discontinuity immediately 3′ to the region encoding the abundantly expressed *gag* gene and 5′ to the regions encoding reverse transcriptase and integrase enzymes. This discontinuity is either a frameshift, in the case of Rous sarcoma and HIV 1 retroviruses and the Ty retrotransposon of yeast (9−13), or a stop codon for murine and feline leukaemia viruses (14,15).

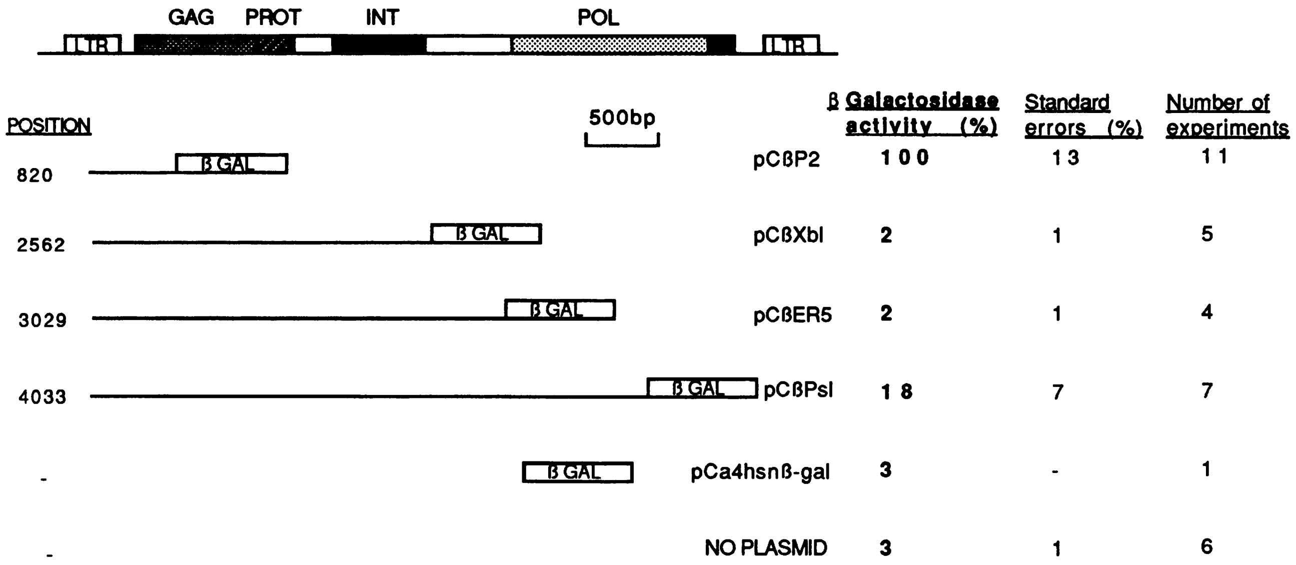
All retroviruses, and almost all retrotransposons, contain such translational discontinuities between their *gag* and *pol/int* regions. The sole exceptions known to date are the *copia* mobile element of *Drosophila* and two closely related elements from plants (16−18). The experiments described here address the question of whether *copia* also expresses lower levels of the analogous proteins to those under-expressed by its distant relatives, and if so by what mechanism. We show here that the protein expression of the *pol/int* region of *copia* is indeed far lower than that of the *gag* region but the mechanism whereby this is achieved differs from the retroviral paradigm. Instead of a single RNA, *copia* uses its subgenomic 2kb RNA to specify the large majority of its *gag* products, while the full length *copia* RNA, containing the *gag*, *pol* and *int* regions, is expressed as protein at a far lower level.

MATERIALS AND METHODS

Construction of Recombinants

All constructs are based upon the *copia* element of the *white-apricot* allele (16) and the vector, plus β galactosidase open

* To whom correspondence should be addressed



POSITION		β Galactosidase activity (%)	Standard errors (%)	Number of experiments
820	pCβP2	100	13	11
2562	pCβXbl	2	1	5
3029	pCβER5	2	1	4
4033	pCβPsl	18	7	7
-	pCa4hsnβ-gal	3	-	1
-	NO PLASMID	3	1	6

FIGURE 1. β Galactosidase Expression From *copia* Fusion Constructs. The relative positions of the β Galactosidase protein coding region in the various constructs used in this study are shown, relative to the approximate positions of the putative *copia* gene products, deduced from the DNA sequence (16). The fusions are to Pvu ll, (pCβP2), Xba l (pCβXbl), Eco RV (pCβER5) and Pst l (pCβPsl) sites at the nucleotide positions noted. β galactosidase activities are expressed as percentages relative to the activity of pCβP2 and the numbers of individual experiments and standard errors of the quoted values are shown.

reading frame, are from pCa4hsnβ-gal (19) or a derivative lacking the small Bam HI fragment in the polylinker (pCa4hsnβ-galΔBam). The *copia* cDNA clones derived from the 2kb RNA were isolated from the Q (late pupal) phage library of Tom Kornberg. pCβP2 fuses the *copia* open reading frame at nucleotide 822, a Pvu II site, to the Sal I site of pCa4hsnβ-galΔBam which was made blunt ended with S1 nuclease. pCβXb1 joins position 2566, an Xba I site made blunt ended by filling in with Klenow polymerase, to Sal I/Klenow-treated pCa4hsnβ-galΔBam. pCβER5 ligates position 3031, an Eco RV site, to the same vector preparation as for pCβXb. pCβPs1 joins position 4037, a Pst I site treated with S1 nuclease, to the same vector preparation as pCβP2. pCβ[2+5] fuses the β galactosidase open reading frame to position 4658 of *copia*, at which point a Bam Hl site has been created by *in vitro* mutagenesis, replacing the sequence TCGAATGCT with TCGGATCCT. The fusion point, between *copia* and β galactosidase, of pCβ3[2] is identical to that of pCβ[2+5] but pCβ[2] lacks nucleotides 1606–4554, because it is derived from the cDNA clone (nucleotides 1571 to 4658 derive from this clone). Additionally, it carries the genomic *copia* sequence from the *white apricot copia* element from position 1 to 1571 (an Lsp I site). pCβ3[5] is identical to pCβ3[2+5] but carries a mutation in its 3′ splice acceptor site which changes it from ATTCCTACAG to ATGTCGACGG. For pCβ3[5Δ], the β galactosidase open reading frame is joined to the Sal I site (GTCGAC) in the mutated splice acceptor site. The preservation of the reading frames across all the above fusion breakpoints and sequences of the mutations, and the entire sequences of the *in vitro* mutagenized regions described above, were confirmed by sequence analysis (20,21)

Cell Transfection and Expression Analysis

Transfections of DH 33 *Drosophila* cells (22) were as described previously (23) but contained from 7μg to 20μg of β galactosidase fusion plasmid and 5 μg pKSCopCAT as an internal control for transfection efficiency. These quantities were determined as non-saturating (data not shown) and the amounts used for each batch of cells were self-consistent and controlled for by the inclusion of two or more pCβP2 transfections in each batch. pKSCopCAT contains a *copia*-CAT fusion construct (23). β galactosidase assays were by the method of Miller (24) and CAT enzyme assays were as described by Gorman *et al*, (25). The results from each batch of transfections were corrected for the individual transfection efficiencies, as controlled internally by CAT enzyme expression, then normalized to the β galactosidase expression of pCβP2 in that batch (arbitrarily denoted as 100%). Each quoted β galactosidase value is an average of two individual assays. The numbers of individual transfected dishes used and the standard errors of each mean value are shown in the Figures.

Proteins were isolated from a set of transfection experiments by immunoprecipitation with rabbit anti-β galactosidase antibodies (obtained from Jackson Laboratories). The relative amounts of protein extract used per lane were corrected for the individual transfection efficiency of each sample, as assayed by CAT activity of the internal control plasmid. The immunoprecipitates were electrophoresed on an 8% denaturing polyacrylamide gel and electrophoretically blotted to nitrocellulose filter. β galactosidase-containing proteins were visualized using mouse anti-β galactosidase antisera (Sigma) followed by alkaline phosphatase-coupled rabbit anti-mouse antibody (Sigma).

RESULTS

To address the question of whether the expression levels of the *copia* pol/int polypeptide gene products are significantly lower than that of the *gag* polypeptides, we needed an assay for protein expression from different regions of *copia*. To achieve this, we used a similar approach to that employed for the *Ty* retrotransposon, namely the use of gene fusions (11). The *lac Z* gene was fused in frame to various positions along the single long open reading frame of the *copia* element (Figure 1). These constructs were then introduced into DH 33 cultured *Drosophila* cells by calcium phosphate-mediated transfection. The expression

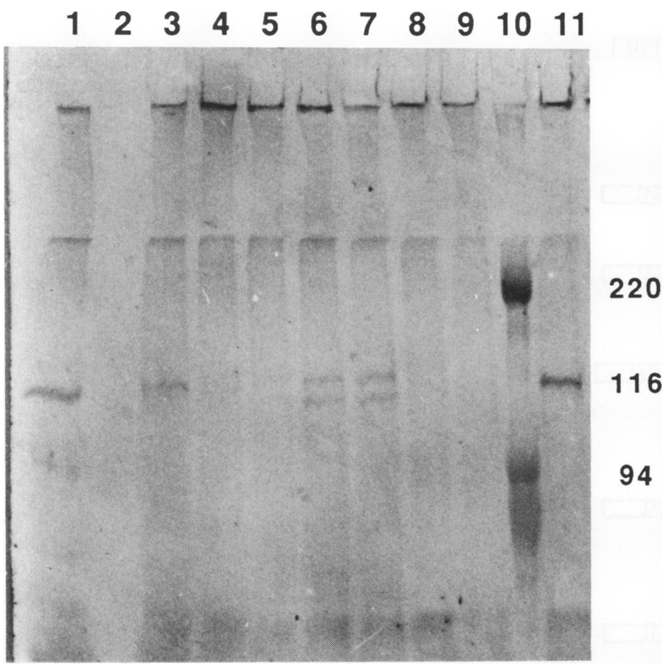


FIGURE 2. Levels of β Galactosidase Protein Correspond to the Respective Levels of Enzyme Activity for the *copia* Constructs. A Western blot of immunoprecipitated β galactosidase-containing proteins isolated from cells transfected with the fusion constructs described in Figures 1 and 3 was probed with anti-β galactosidase antibody. Lane 1; 30ng β galactosidase. Lane 2; Blank. Lanes 3 and 11; pCβP2. Lane 4; pCβXb1. Lane 5 pCβPs1. Lane 6; pCβ[2]. Lane 7; pCβ[2+5]. Lane 8; pCβ[5]. Lane 9; CAT plasmid alone. Lane 10; Molecular weight markers (mobilities in kiloDaltons are shown).

of β galactosidase enzyme activity from the constructs was measured in extracts of these cells 72 hr later (23).

Figure 1 shows that fusions to the *copia int* and *pol* gene regions result in much lower levels of β galactosidase activity than a fusion to the *gag* gene (pCβP2). Fusions to positions 2566 and 3031 yield β galactosidase activities that are not significantly different from the background level of the assay, which is approximately 3% of the value for pCβP2. The activity of the fusion to position 4040 is measurably higher but still at least a factor of five-fold lower than pCβP2.

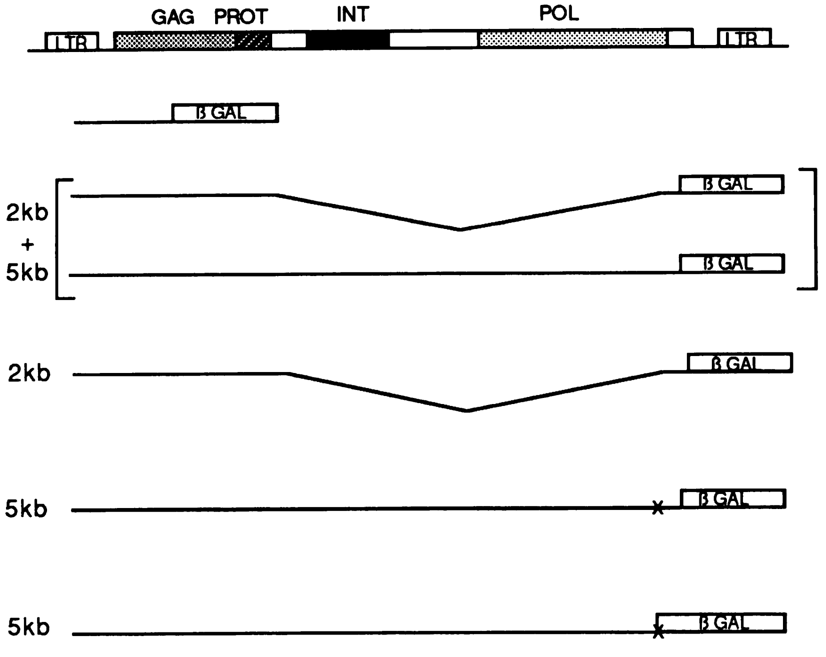
To ensure that the lower activities of the *int* and *pol* fusions was not due to differences in enzyme activity of the constructs resulting from their possessing different amounts of *copia* protein attached to their amino termini, we measured the amounts of β galactosidase protein immunologically. Figure 2 shows that the *gag*-β galactosidase protein fusion construct derived from pCβP2 (lanes 3 and 11) is present in much greater amounts than the fusions to the *int/pol* region (pCβXb1and pCβPs1), which are almost undetectable by this method (lanes 4 and 5 respectively).

We next addressed the question of how this effect is achieved. *copia* encodes an extra RNA in addition to the full length species (26). This subgenomic 2 kilobase (kb) RNA is present in approximately similar amounts to the full length RNA in cultured cells (27). Furthermore, it has the same 5′ terminus and at least 800 bases of 5′-proximal sequence as the full length RNA but lacks the *pol* region (27–28). It therefore seemed likely that it might encode the *gag* gene products exclusively.

Mount and Rubin had previously suggested, on the basis of sequence analysis, that the 2kb *copia* RNA might be spliced (16). However, its detailed structure was unknown. We therefore isolated, from a pupal cDNA library, clones fulfilling hybridization criteria of the presence of *gag* region and absence of *int* and *pol* sequences as predicted by the mapping data of Schwartz *et al* (28). Six clones were isolated, two of which were apparently derived from a spliced RNA. Sequence analysis of these clones showed a splice donor site at nucleotide 1605 (the donor site predicted by Mount and Rubin, reference 16) and an acceptor site at position 4555. The 5′ exon of the spliced RNA encodes the entire *gag* region, plus the region homologous to the putative protease. The 3′ exon contains the last 34 amino acids of the *copia* long open reading frame, the splice having preserved the translational reading frame. The splice junctions are a good match to the splice site consensus sequences (29). Recently Miller *et al* and Yoshioka *et al* have reported the isolation and sequence analysis of similar clones of the same RNA (30, 31). Incidentally, one of our clones derived from an RNA with a poly A tail attached to nucleotide 5083 as opposed to the previously observed 5092 (30,32). This is not too surprising because *copia* does not possess a perfect AATAAA polyadenylation sequence (16).

The characterization of a subgenomic 2 kilobase RNA with coding potential for the *gag* and protease regions suggests a mechanism for the lower levels of expression of *int* and *pol* gene products. The full length RNA must account for the *int* and *pol* products (it is the only known *copia* RNA capable of encoding these proteins) and the 2kb RNA might be responsible for the majority of the *gag* protein production. To test this possibility, we fused the β galactosidase protein coding region to a position a short distance downstream of the splice acceptor site, again in the same reading frame as the *copia* long open reading frame and assayed this construct for β galactosidase expression in cultured cells as before (Figure 3). Such a construct (pCβ[2+5]) should yield fusion proteins from both spliced and unspliced RNAs and would therefore, by the above hypothesis, generate high level β galactosidase expression. This is in fact the case, although the absolute levels are approximately half that of pCβP2. This reduced expression is probably not due to translational effects of the extra 300 amino acids of *copia* on the β galactosidase because the levels of protein, as assayed immunologically, are similarly lowered (Figure 2 lane 7). Interestingly, there are two bands in this lane, the smaller of which is the approximate size of authentic β galactosidase (lane 1), suggesting that this fusion is being cleaved in the cells. In support of this suggestion, the polyprotein encoded by the 2kb *copia* mRNA has been shown to cleave itself *in vitro* (32, our unpublished observations)

Next we designed constructs to test the relative protein expression levels of *copia's* 2kb and 5kb RNAs *in vivo*. The 2kb construct (pCβ[2]) comprises a reconstructed complete cDNA copy of the 2kb RNA with a complete *copia* LTR, containing the *copia* promoter, replacing the actual 5′ RNA end. The β galactosidase open reading frame was fused to the same position in *copia* as for pCβ[2+5]. Because the 2kb RNA's intron has been removed at the DNA level, this construct can only produce the spliced fusion RNA. To create a construct which is only capable of producing exclusively unspliced RNA (pCβ[5]), we destroyed the 3′ splice acceptor site of the 2kb RNA. Results from other laboratories have shown that removal of the splice acceptor sequence abolishes splicing (33,34), unless there is an alternative acceptor available. The β galactosidase protein coding region was fused to the same position in the *copia* open reading frame as the other two constructs. A final construct, fusing the β galactosidase open reading frame to the mutated splice acceptor



	β Galactosidase activity (%)	Standard errors (%)	Number of experiments
pCβP2	100	12.5	11
pCβ[2+5]	56	17	9
pCβ[2]	75	28	8
pCβ[5]	0.1	0.9	8
pCβ[5Δ]	6	0.9	3

FIGURE 3. Fusion Constructs based upon the *copia* 2kb RNA are expressed more efficiently than constructs derived from the 5kb *copia* RNA in cultured cells. β galactosidase expression levels from fusion constructs designed to express β galactosidase from either the *copia* 2kb RNA (pCβ[2]), 5kb RNA (pCβ[5] and pCβ[5Δ]) or both 2kb and 5kb RNAs (pCβ[2+5]) were measured. β galactosidase activities are expressed as percentages relative to the activity of pCβP2 and the numbers of individual experiments and standard errors of the quoted values are shown.

site at position 4555 (pCβ[5Δ]), was made, in case the mutated acceptor sequence had fortuitously created a spurious acceptor site, or the 3′ exon contained a cryptic splice acceptor site.

The expression of these constructs was analysed in the same way as before. Figure 3 shows that pCβ[2], which can only produce spliced *copia*- β galactosidase RNA, gives high level expression, whereas both pCβ[5] and pCβ[5Δ], which can only produce unspliced fusion RNAs, give low levels of β galactosidase. Again, the levels of enzyme activity are broadly reflected by the protein levels (Figure 2, lanes 6–8). Therefore, in the cells used in this study, fusion constructs derived from 2kb RNA are expressed much more efficiently than the equivalent constructs that can only synthesize 5kb-based fusion RNAs.

DISCUSSION

These data show that the *gag* gene products of *copia* are produced in far greater amounts than the *pol/int* products and that this differential expression is achieved by the provision of the separate subgenomic 2kb RNA which encodes *gag* gene products exclusively. This RNA is expressed as protein with a greater efficiency than the unspliced 5kb RNA. Thus, *copia* achieves the same result as retroviruses and the Ty retrotransposon, but by a fundamentally different mechanism.

How is the 2kb *copia* mRNA expressed more efficiently than the 5kb species? At present we do not know the answer to this question. It is very unlikely that there are significantly more 2kb-derived RNAs than 5kb-derived RNAs in our experiments, since the endogenous 2kb and 5kb RNAs are expressed at very high and equivalent levels in cultured cells (27,28).

Another possibility is that the different *copia* polypeptides which are fused to the N-terminus of the β galactosidase donate varying degrees of stability to the fusion proteins. We cannot exclude this possibility but such a phenomenon has not been previously reported, despite the widespread use of β galactosidase fusions (35). The only indication of protein cleavage in our studies leaves an apparently stable, and more or less intact, β galactosidase which we suggest is generated by autocatalytic processing by the *copia*-encoded protease (31).

We therefore favour the alternative explanation that the higher expression of 2kb RNA-derived fusion RNA is due to an inherently higher translation efficiency of *copia* 2kb RNA This hypothesis is supported by our previous observation that *copia* 2kb RNA, isolated from *Drosophila* cultured cells, is also translated more efficiently than is the 5kb RNA in the heterologous rabbit reticulocyte system (36). Such a difference in expressibility would be interesting, because the two RNAs are identical for the first 1.5kb of their sequence (27,30, this work) and the translational initiation codon for the putative *copia* polyprotein lies approximately 0.3kb from the 5′ end of the RNAs (16). Because the only difference between the two RNAs is the presence in the 5kb *copia* RNA of nucleotides 1606 to 4554, a sequence which is responsible for this effect must lie in these 2949bp. We are presently investigating this further.

We were unable to detect fusion protein expression from four of the five *pol/int* fusions. This does not mean that there is none produced, only that the amounts fall below our detection level (approximately 3% of the value for pCβP2). This implies that the *copia pol* and *int* proteins are perhaps synthesized in lower amounts, relative to the *gag* proteins, than are the equivalent retroviral products. The detectable protein expression level from the fifth *pol/int* fusion, pCβPs1, is puzzling. Could there be a cryptic splice acceptor site which is specifically activated with this construct? Such a site cannot lie entirely within the the *copia* sequence, because it would also be present in both pCβ[5] and pCβ[5Δ], which produce no detectable β galactosidase. Therefore, the only possible site is at the fusion junction. This sequence is GATTTAACTACTCTAG/AGGATC. This bears some resemblance to the consensus acceptor site (29) and a splice at the position noted (/) would preserve the open reading frame. We therefore favour this explanation, though further studies are required.

ACKNOWLEDGEMENTS

We thank Jane Forrester for excellent technical assistance. We are also grateful to Luke Alphey for useful discussions, to Steve Hyde for advice on *in vitro* mutagenesis, to Tom Kornberg for his cDNA library and to Alistair Murchie and David Lilley for oligonucleotide synthesis facilities. This work was supported by the Medical Research Council.

REFERENCES

1. Varmus, H.E. (1983) p411–503. In J.A.Shapiro (ed) Mobile Genetic Elements. Academic Press Inc., Orlando, Fla USA.
2. Varmus, H.E. and Brown, P. (1989) in Berg, D.E. and Howe, M.M. (eds) Mobile DNA published by American Society for Microbiology, Washington USA; pp53–108
3. Finnegan, D.J. (1985) International Review of Cytology, **93**, 281–326.
4. Dickson, C., Eisenman, R., Fan, H., Hunter, E. and Teich, N. (1984) In Weiss, R., Teich, N., Varmus, H. and Coffin, J. (eds) RNA Tumor Viruses. Published by Cold Spring Harbor Laboratory, New York, pp513–648.
5. Dayton, A.I., Sodrowski, J.G., Rosen, C.A., Goh, W.C. and Haseltine, W.A. (1986) Cell **44**, 941–947.
6. Strebel, K., Daugherty, D., Clouse, K., Cohen, D., Folks, T. and Martin, M. (1987) Nature **328**, 728–730.
7. Terwilliger, E., Burghoff, R., Sia, R., Sodrowski, J., Haseltine, W. and Rosen, C. (1988) J. Virol. **62**, 655–658.
8. Cullen, B.R. and Green, W.C. (1989) Cell **58**, 423–426.
9. Jacks, T.and Varmus, H.E. (1985) Science **230**, 1237–1242.
10. Jacks, T., Power, M.D., Masiarz, F.R., Luciw, P.A., Barr, P.J. and Varmus, H.E. (1988) Nature **331**, 280–283.
11. Mellor, J., Fulton, A.M., Dobson, M.J., Roberts, N.A., Wilson, W., Kingsman, S.M. and Kingsman, A.J. (1985) Nature **313**, 243–246.
12. Wilson, W., Malim, M.H., Mellor, J., Kingsman, A.J. and Kingsman, S.M. (1986) Nucleic Acids Res **14**, 7001–7016.
13. Clare, J.J., Belcourt, M. and Farabaugh, P.J. (1988) Proc Natl Acad. Sci. USA **85**, 6816–6820.
14. Yoshinaka, Y., Katoh, I., Copeland, T.D. and Oroszlan, S. (1985a) J.Virol. **55**, 870–873.
15. Yoshinaka, Y., Katoh, I., Copeland, T.D. and Oroszlan, S. (1985a) Proc Natl Acad Sci USA **82**, 1618–1622.
16. Mount, S.M. and Rubin, G.M. (1985) Mol. Cell. Biol. **5**, 1630–1638.
17. Voytas, D.F. and Asubel, F.M. (1988) Nature **336**, 242–244
18. Grandbastein, M-A., Spielman, A. and Caboche, M. (1989) Nature **337**, 376–380.
19. Martin, M., Meng, Y. B. and Chia, W. (1989) Mol Gen. Genet. In press
20. Maxam, A. M. and Gilbert, W. (1980) Methods in Enzymology **65**, 499–559.
21. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA **74**, 5463–5467.
22. Sondermeijer, P.J.A., Derksen, J.W.M. and Lubsen, N.H. (1980) In Vitro **16**, 913–914
23. Sneddon, A. and Flavell, A.J. (1989) Nucleic Acids Res. **17**, 4025–4035.
24. Miller,J.H. (1972) pp352–355 In 'Experiments in Molecular Genetics'. Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.
25. Gorman, C.M., Moffat, L.F. and Howard., B.H. (1982) Mol. Cell. Biol. **2**, 1044–1051.
26. Carlson, M. and Brutlag,D. (1979) Proc. Natl. Acad Sci. USA **75**, 5898–5902.
27. Flavell, A.J., Levis,R., Simon.M.A. and Rubin., G.M. (1981) Nucleic Acids Res. **9**, 6279–6291.
28. Schwartz, H.E., Lockett, T.J. and Young, M.W. (1982) J. Mol. Biol. **157**, 49–58.
29. Mount, S.M. (1982) Nucleic Acids Res **10**, 459–472.
30. Miller, K., Rosenbaum, J., Zbrzezna, V. and Pogo, A.O. (1989) Nucleic Acids Res **17**, 2134.
31. Katsuji, Y., Honma, H., Zushi, M., Kondo, S., Togashi, S., Miyake, T. and Shiba, T. (1990) EMBO J. **9**, 535–541.
32. Emori, Y., Shiba, T., Kanaya, S., Inouye, S., Yuki, S. and Saigo, K. (1985)Nature **315**, 773–776.
33. Wieringa, B., Hofer, E. and Weissman, C. (1984) Cell **37**, 915–922.
34. Aebi, M., Hornig, H., Padgett, R.A., Reiser, J. and Weissman, C. (1986) Cell **47**, 555–565.
35. Silhavy, T.J., Berman, M.L. and Enquist, L.W. (1984) pp18–23 in 'Experiments with Gene Fusions' Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.
36. Flavell, A.J., Ruby,S.W., Toole, J.J., Roberts, B.E. and Rubin., G.M. (1980) Proc. Natl. Acad. Sci. USA **77**, 7107–7111.