POLYOMA VIRUS: POLYADENYLATION AND RECOMBINATION

Christopher John Norbury

Thesis presented in partial fulfilment of

the degree of Doctor of Philosophy

to the University of London

Imperial Cancer Research Fund
Laboratories,
Lincoln's Inn Fields,
London, WC2A 3PX.

Department of Biochemistry,
Imperial College of Science
and Technology,
London, SW7 2AZ.

September, 1987.

For Gail, for Coventry and for the hell of it.

ABSTRACT

The polyoma virus (Py) early region is a complex transcription unit whose primary nuclear transcripts are differentially spliced to produce at least three mRNAs encoding the large, middle and small T-antigens (T-Ags). The position of an alternative polyadenylation (poly(A)) site at the 3' end of the polyomavirus (Py) middle T-antigen coding sequences suggests the possibility of a functional role for this site in early gene regulation. The fine structure of this alternative poly(A) site was determined by cDNA sequence and 3' S1 analyses. Cleavage/polyadenylation was found to be heterogenous occurring at multiple CA dinucleotides downstream from the AATAAA signal sequence. About 50% of the alternative polyadenylation takes place upstream from the middle T-Ag stop codon. In addition, the pattern of splicing of transcripts using the alternative poly(A) site was found to differ from those using the major poly(A) site at the end of the early region. The ratio of the small and middle T-Ag splices to the large T-Ag splice for the alternative polyadenylated mRNAs is about 2.5 times that found for mRNAs using the major Poly(A) site. The altered splicing pattern and 3' end heterogeneity of the alternative polyadenylated mRNAs would result in preferential translation of small T-Ag (to a greater degree) and middle T-Ag over large T-Ag at later times in the Py lytic cycle.

The pattern of polyadenylation in the Py-transformed cell line 53-Rat is unusual in that the majority of viral transcripts are polyadenylated at the alternative viral poly(A) site. The extent to which downstream sequences are responsible for this phenomenon was investigated using a novel transient expression assay. Sequences from a 1 kilobase region of flanking DNA were shown to impair utilisation of an efficient poly(A) site downstream, while enhancing polyadenylation at a less efficient site located upstream from these sequences. Other lines of evidence suggested that cellular sequences flanking the site of Py integration are the principle determinant of the pattern of viral polyadenylation in transformed cells.

The molecular events involved in Py integration were investigated by determining the arrangement of viral and cellular sequences in two transformed cell lines. The results served both to confirm and extend the conclusions of earlier studies. For one line, 82-Rat, the discovery of two formerly contiguous viral sequences at different recombinant joins made possible the formulation of a model for viral integration.

ACKNOWLEDGEMENTS

THANKS

-to Mike Fried, for taking so much abuse, often in the spirit in which it was intended.

-to Moira Read, for showing me how to conduct myself properly when in the presence of cultured individuals.

-to Brendan Davies and Claudio Passananti, for many stimulating and helpful discussions.

-to the many others who have made my time at I.C.R.F. so interesting.

-to Peter Rigby and David Glover, for supervising my studies at Imperial College.

-and to Gail, for making it all worthwhile.

ABBREVIATIONS USED

Py	....murine polyoma virus
SV40	....simian virus 40 (vacuolating virus)
HaPV	....hamster papova virus
T-Ag	....tumour antigen
VP	....virion protein
Ad	....adenovirus
Ig	....immunoglobulin
nt., nts	....nucleotide, nucleotides
bp	....base pair(s)
kb	....kilobase(s)
kDa	....kilodalton(s)
pfu	....plaque forming units
FCS	....fœtal calf serum
MOPS	....3-[N-Morpholino]propanesulphonic acid
PIPES	....Piperazine-N, N'-bis[2-ethanesulphonic acid]
Tris	....Tris[hydroxymethyl]aminomethane
SDS	....sodium dodecyl sulphate
EDTA	....diaminoethane tetraacetic acid (disodium salt)
BSA	....bovine serum albumin
DTT	....dithiothreitol
mins	....minutes
hrs	....hours
RT	....room temperature

CONTENTS

FIGURES AND TABLES

FIGURES

CHAPTER ONE: INTRODUCTION

This chapter contains a summary of polyomavirus biology and molecular genetics. The early sections present an account of the viral lytic cycle, which in many respects amounts to a virus-orientated view of eukaryotic gene expression and its control. The emphasis here is on recent findings in the field, and, in describing essentially cellular processes of which the viruses take advantage, frequent reference is made to other, cellular systems. In the interests of completeness, accounts of viral splicing and polyadenylation are relatively detailed, and are supplemented by reviews of the analogous cellular processes. The subsequent sections involve polyomavirus-mediated cell transformation and, finally, a review of non-homologous recombination in eukaryotic cells.

BIOLOGY OF THE POLYOMAVIRUSES

During the past thirty years members of the *papovaviridae* family (Melnick *et al.*, 1974) have been the deserving recipients of an enormous amount of scientific attention. The more recent past has seen an explosion of interest in the papillomaviruses of the genus *papovavirus* A, largely due to the apparent involvement of human papillomaviruses in the genesis of cervical carcinomas. Nonetheless, the majority of papovavirus research has involved the polyomaviruses (genus *papovavirus* B), in particular simian vacuolating virus (SV40) of rhesus monkeys and, to a lesser extent, polyoma virus (Py) of mice.

The most exciting aspect of polyomavirus biology is the fact that, despite their very small genome size, viruses of this genus are able to cause the formation of tumours in susceptible animals (Gross, 1953; Eddy *et al.*, 1962; Girardi *et al.*, 1962). Since the number of potential interactions of viral components and products with the host was presumed to be rather small, it was reasoned that elucidation of the mechanism of tumorigenesis might be a realistic aim. The discovery of cell culture systems in which viral growth (Sweet and Hilleman, 1960; Stewart *et al.*, 1957) and transformation (Vogt and Dulbecco, 1960; Todaro and Green, 1964) can be studied has greatly facilitated polyomavirus research. While their oncogenic potential continues to be a source of great interest, the viruses have also proved to be invaluable model systems for the study of transcriptional control both in *cis* and in *trans*, RNA processing by differential splicing and polyadenylation, DNA replication and macromolecular assembly (reviewed in Tooze, 1981 and Salzman, 1986).

Polyomaviruses in the wild.

It was the tumorigenicity of polyomaviruses which generated much of the early interest in this field; this, indeed, was the property by which Py was first identified (Gross, 1953). With so much emphasis having been placed on cellular transformation,

especially since the identification of the transforming gene products of SV40 (Martin and Chou, 1975; Brugge and Butel, 1975) and Py (Treisman *et al.*, 1981b), it is important to remember that neither virus appears to cause cancer in the wild. In fact, it is striking that of the eleven members of the polyomavirus genus so far described, only one, the hamster papovavirus (HaPV), is implicated in tumour formation in its natural environment (Table 1.1). The human JC virus (*Papovavirus* B, type 6) is an opportunistic pathogen causing the destructive and usually fatal brain disease progressive multifocal leukoencephalopathy (PML) (Padgett *et al.*, 1971). Similarly, the recently discovered Budgerigar Fledgling Disease Virus (BFDV) is associated with a contagious acute disease resulting in the death of affected birds (Muller and Nitschke, 1986). In contrast, neither SV40 nor Py appears to be pathogenic to any degree in their respective natural host, rhesus monkey or mouse. Thus, where members of the *papovavirus* B genus are associated in nature with disease, these diseases are usually cytopathic, rather than neoplasatic. In particular, neither JC virus nor BK virus, the other human virus of the genus, has been shown to be involved in human neoplasia, despite extensive searches for viral T-antigens (T-Ags) or DNA sequences in human tumours, and for anti-T-Ag antibody in sera from cancer patients (reviewed by Takemoto, 1980).

Polyomaviruses in the laboratory.

The general lack of involvement of polyomaviruses in tumorigenesis in the wild is in marked contrast to the behaviour of these viruses when artificially introduced, in relatively large amounts, into homologous or heterologous host animals. Table 1.1 summarises the available data relating to tumour formation following injection of purified virus preparations at a variety of sites in a number of different hosts. In general, tumours generated in this manner are most commonly found in immunologically immature or immunodeficient animals. The failure of immunologically competent animals to develop tumours is widely considered to be a measure of the effectiveness of the immune rejection

Polyomavirus Species	Natural Host	Systems in which Tumorigenic	Tumours Formed	Reference
SV40	Rhesus monkey	Hamster: newborn	Fibrosarcomas at site of injection	Eddy et al., 1962.
		adult	Leukaemias, lymphomas, osteosarcomas, reticulum cell sarcomas (high I.V.dose of virus required)	Diamandopoulos, 1978.
		Mouse (with low cytotoxic T cell response to TSTA)	Lymphomas	Abramczuk et al., 1984
Py	Mouse	Newborn mouse, hamster, rat	Multiple tumours of salivary glands, thymus, ovaries, mammary glands and adrenals	Gross, 1953, Eddy et al., 1958
		Adult hamster	Fibrosarcomas at site of injection	Defendi, 1960
HaPV	Hamster	Hamster: newborn	Lymphomas, leukaemias	Graffi et al., 1970,
		young	Naturally occurring, spontaneous hair follicle tumours	Graffi et al., 1967
JC	Human	Newborn hamster, adult owl monkey, squirrel monkey	A wide variety of CNS tumours, but no subcutaneous tumours	Walker et al., 1973, London et al., 1978, Miller et al., 1984
BK	Human	Newborn hamster, mouse, rat	A wide variety, with a degree of virus strain-specificity	Shah et al., 1975, Uchida et al., 1979

Table 1.1. Tumorigenicity of the extensively studied polyomaviruses (for discussion see text)

of cells bearing the tumour-specific transplantation antigen (TSTA), which appears on the surface of cells infected or transformed by these viruses. This view is supported by the results of injection of SV40 into different mouse strains (Abramczuk *et al.*, 1984). These studies showed that high viral tumorigenicity in H-2^d mice correlated with the host's reduced ability to mount a cytotoxic T-lymphocyte response to the SV40 TSTA, as compared with H-2^k mice, which did not develop tumours. In the light of these results, one could postulate that the subcutaneous fibrosarcomas which develop, after a long latency period, at the site of injection of Py into adult hamsters (Defendi, 1960) indicate that very high local concentrations of virus are occasionally sufficient to circumvent TSTA-mediated rejection. Presumably, it is the unnaturally high level of virus used which accounts for the ability of SV40 to induce a variety of tumours after intravenous injection into adult hamsters (Diamandopoulos, 1978), but it is not known how the spontaneous hair follicle tumours caused naturally by HaPV (Graffi *et al.*, 1967) are able to evade the host's immune surveillance. It is a sobering thought that SV40 has been inadvertently injected into an enormous number of people as a contaminant of poliovirus vaccines produced in rhesus monkeys (Sweet and Hilleman, 1960). Fortunately, the result of this unintentional *in vivo* experiment seems to be that SV40 is not oncogenic in humans, since no tumours in the individuals involved have been shown to involve the virus as a causative agent.

Py AND SV40: THE LYTIC CYCLE

The vast majority of what we know about the course of polyomavirus productive infection has come from studies using Py and SV40, and I shall restrict my review of this aspect of polyomavirus biology to these two viruses. Since the subject has been reviewed extensively in the recent past (Tooze, 1981), this account is largely limited to aspects which appear relavent to the questions investigated in later chapters.

Productive infection occurs only in cells which permit viral DNA replication: monkey cells for SV40 and mouse cells for Py. The viruses exhibit biphasic gene expression, with late gene products with virion structural roles produced only after DNA replication and early gene products, whose roles are essentially regulatory, produced throughout the cycle. Synthesis of progeny particles is followed by cell lysis.

Virion structure.

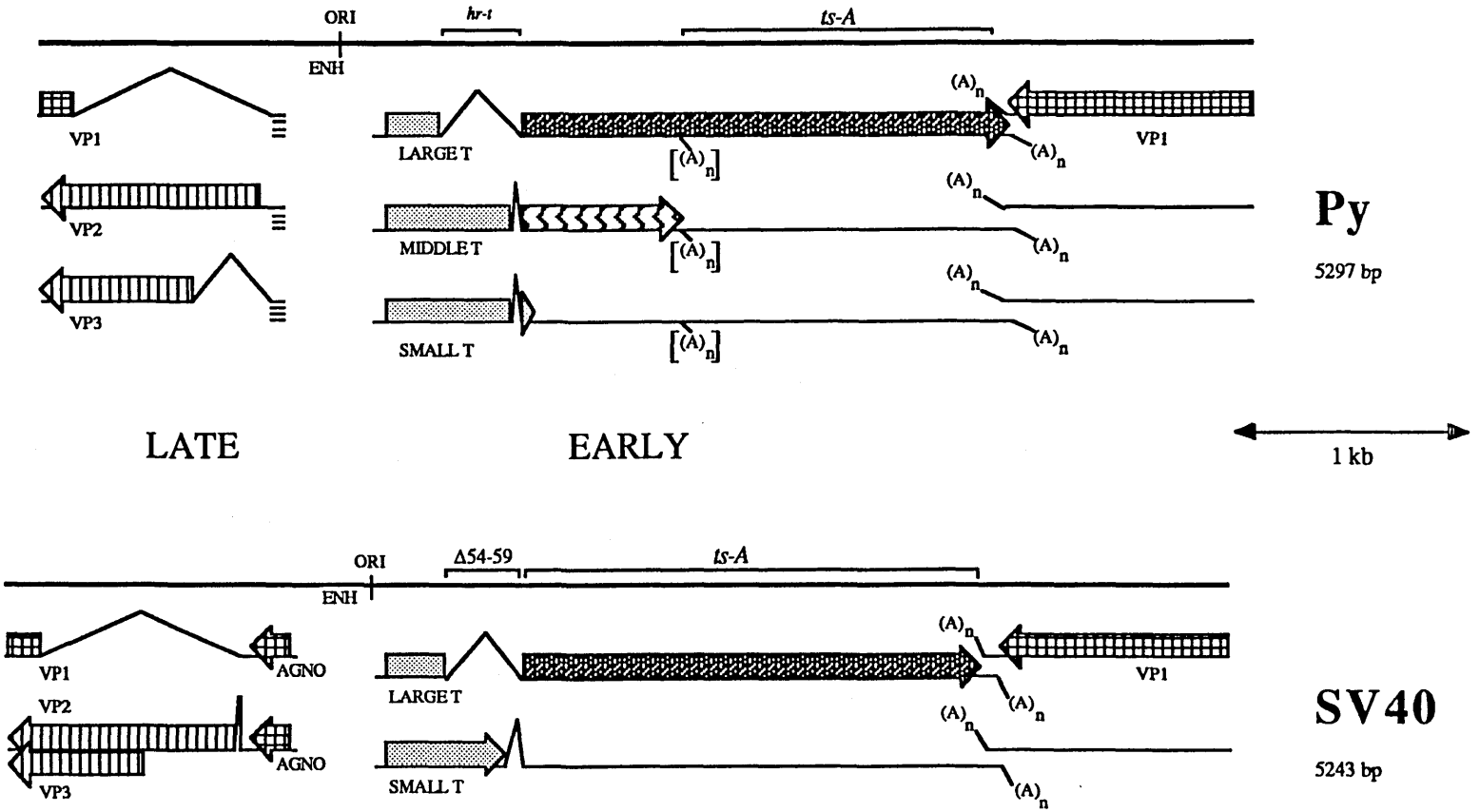
The virions of both SV40 and Py are icosahedral particles some 45nm in diameter containing a single supercoiled, covalently closed, circular double-stranded DNA viral genome of about 5.3 kb. The viral DNA is complexed with the cellular histones H2A, H2B, H3 and H4 in a chromatin-like configuration, and this is surrounded by the viral capsid, which consists of virally encoded VP1 and, to a lesser extent, VP2 and VP3. VP1 is a 45-47 kDa protein which is post-translationally modified to give a family of six species (A-F) of differing pIs. The roles of SV40 and Py proteins VP2 and VP3 are not well understood, although one possibility is that together they form the 12 capsomeres which are situated at the vertices of the icosahedral virion (Brady *et al.*, 1978). The significance of the recent finding that VP2 is myristilated (Streuli *et al.*, 1987) is not known, but this post-translational modification offers the possibility of generating lipophilic regions at the virion surface.

Genome organisation.

The Py and SV40 genomes display a similarity of genetic organisation which is held in common with other polyomaviruses (Fig. 1.1). The most striking feature of this organisation is the absence of nonfunctional sequences. Not only is more than 90% of each genome accounted for by protein coding sequences, but extensive use is also made of overlapping reading frames, particularly in Py. This remarkable genetic economy is presumably evidence of evolutionary selection for minimum genomic size, which would

Figure 1.1.

Organisation of protein-coding and regulatory sequences in the Py and SV40 genomes. Each genome is represented in a linearised form by the heavy line at the top of each half of the figure, and the genomes are aligned to demonstrate structural and functional homologies. The early and late transcription units are separated by the central control region containing the origin of replication and transcriptional enhancer sequences. The map positions of mutations which result in thermolabile large T-Ag proteins (*tsA*) and of deletion mutations which abolish Py transforming activity (*hr-t*) or severely impair SV40 transformation (Δ54-59) are indicated above. Below each genome, the structure of the predominant early and late mRNAs (thin lines) are shown, with chevrons denoting differential splicing and $(A)_n$ representing the 3' poly(A) tracts. The reiterated leader sequence at the 5' ends of Py late mRNAs and the alternative poly(A) site in the middle of the Py early region are also shown. Thick shaded arrows represent the protein coding sequences; utilisation of different open reading frames is indicated by differential shading. The short reading frame near the 5' ends of the SV40 late mRNAs is that of the 61-amino-acid agnoprotein.


ORI
hr-t
ts-A
ENH
VP1
VP2
VP3
LARGE T
MIDDLE T
SMALL T
VP1
Py
5297 bp
LATE
EARLY
1 kb
ORI
Δ54-59
ts-A
ENH
VP1
AGNO
VP2
AGNO
VP3
LARGE T
SMALL T
VP1
SV40
5243 bp

be advantageous in the rapid generation of large numbers of progeny viral minichromosomes late in viral infection.

Py and SV40 early and late region protein coding sequences lie on opposite strands of the viral DNA, and are separated at their 5' ends by a control region, which houses transcriptional enhancer and promoter elements as well as the origin of bidirectional DNA replication. Within each late region, the overall layout of VP1, VP2 and VP3 coding sequences is very similar, but the SV40 late region is distinguished in having an additional short open reading frame at the promoter-proximal end. This encodes the 61-amino-acid DNA-binding agnoprotein (Jay *et al.*, 1981), which may participate in virion assembly (Ng *et al.*, 1985) and/or late viral gene regulation (Alwine, 1982; Hay *et al.*, 1982). Similarly, comparison of the two early regions reveals gross similarities, but the Py early region, like that of HaPV (Delmas *et al.*, 1985) encodes a third protein product, the middle T-Ag.

Attempting to establish the evolutionary relationship between the two types of early region involves considerable conceptual problems. It is difficult to imagine, for example, that the middle T-Ag sequences of Py were acquired by a progenitor SV40-like virus through recombination with a cellular gene, since an intact large T-Ag reading frame, essential for virus growth, could not easily be retained after such an event. A more plausible explanation is that the degeneracy of the genetic code allowed considerable DNA sequence variation in an ancestral SV40-type viral early region, whilst maintaining essential large T-Ag functions. At some stage in evolutionary time the novel middle T-Ag was fortuitously generated, and endowed the Py ancestor with a new tissue tropism, or even host range. If the new biological niche became exploited by the virus, then selective pressure would have been applied resulting in the maintenance of the middle T-Ag reading frame. In this light it is interesting to note that Py host-range mutants have been shown to map to that part of the early region encoding small and middle, but not large T-Ags (Fig. 1.1, Feunteun *et al.*, 1976). Alternatively, the evolutionary relationship might be

reversed, and SV40-like viruses could have arisen from Py-like ancestors through an abolished requirement for middle T-Ag function in particular biological niches. In either case, it may be significant that the Py-type viruses have an alternative early region polyadenylation (poly(A)) site at the 3' end of the middle T-Ag coding sequences (see below and chapters 2 and 3). A possible explanation for the high degree of DNA sequence homology between the alternative poly(A) site of Py and the corresponding region of HaPV is that the AATAAA sequence encodes important amino acids of both the middle and large T-Ag proteins, in overlapping reading frames. This situation may constitute an evolutionary trap, in that the possibilities for divergence at the level of DNA sequence may be very limited by a requirement to conserve the two protein sequences. A more attractive hypothesis is that the alternative poly(A) site has evolved as a means of differential regulation of early gene expression (see chapter 2).

Early events in *papovavirus* B infection of cells in culture.

Within fifteen minutes of infection, Py virions arrive in the host cell nuclei (MacKay and Consigli, 1976), via a specific transport mechanism which appears to involve VP1 species E (Bolen *et al.*, 1981). Subsequent uncoating to expose the viral chromatin may involve association with host proteins of the nuclear matrix (Winston *et al.*, 1980). When the SV40 nucleoprotein complex from infected cells is digested with micrococcal nuclease, the DNA products are multiples of 150 to 200 bp in length, indicating that the viral genome is normally complexed with nucleosomes (Shelton *et al.*, 1978). This interpretation has been confirmed by electron microscopy of SV40 chromatin isolated early in infection, which further suggested that the region of the minichromosome encompassing the transcriptional control sequences is nucleosome-free (Jacobovits *et al.*, 1982). Furthermore, this part of the minichromosome is DNAse hypersensitive for both Py and SV40 (Gerard *et al.*, 1982; Herbomel *et al.*, 1981). For SV40, the freedom from nucleosomes and DNAse hypersensitivity has been shown to depend upon the local DNA

sequence. Thus, when sequences from the enhancer and early promoter are inserted at a novel position in the viral genome, these sequences specify the location of a new nucleosome-free region (Jacobovits *et al.*, 1982; Jongstra *et al.*, 1984). It is not clear that the positioning of nucleosomes is established by the DNA sequence itself. A more likely explanation is that sequence-specific DNA binding proteins exclude nucleosomes from the central control region, since a number of candidate proteins have been identified; these are described in the following sections.

EARLY GENE EXPRESSION AND ITS CONTROL

Regulation of expression of a given gene can be achieved at one or more of a number of different levels in the eukaryotic cell. Thus transcriptional initiation and termination, and RNA splicing and polyadenylation can change in an apparently regulated manner in specific cellular circumstances (reviewed in Nevins, 1983; see also below). Even after formation of the complete mRNA, nuclear-cytoplasmic transport, translational control and mRNA stability could be involved in gene regulation. Despite the complexity of the pathway of mRNA biogenesis, studies on the regulation of gene expression in general, and on SV40 and Py expression in particular, have tended to focus on regulation at the level of transcriptional initiation.

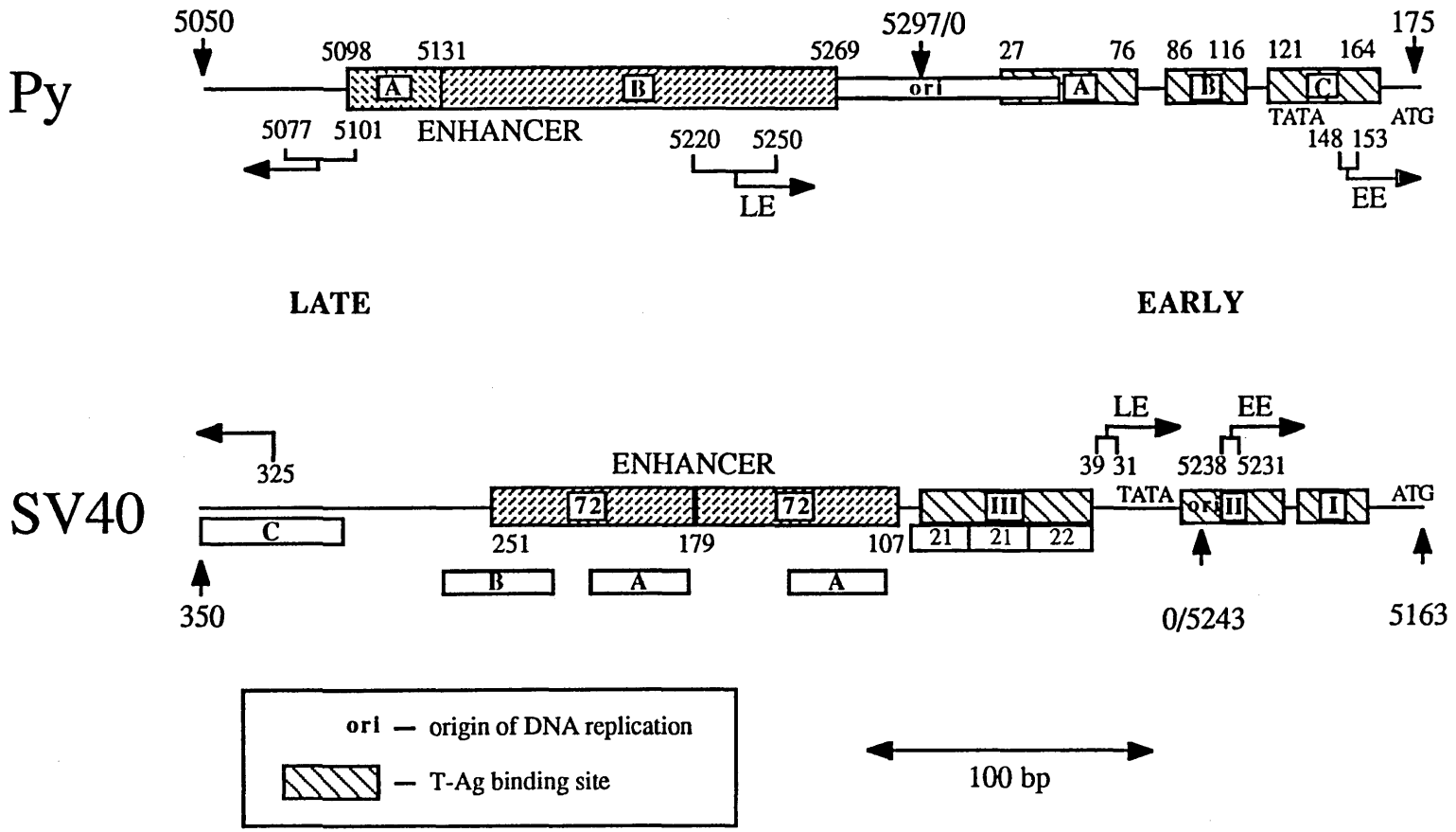
The role of the enhancer.

The unpackaged viral chromatin is accessible to a number of components of the cellular gene expression machinery and it is cellular, rather than viral, factors which appear to bring about the onset of early viral gene expression. Molecular genetic and DNA-protein binding studies have combined in recent years to generate a largely coherent, if incomplete, explanation of early transcription and its regulation. Comparison of the organisation of the central control region of SV40 with that of Py reveals more functional similarities than differences (Fig. 1.2). In particular, both viral genomes have enhancer

elements in this region. The enhancer element is needed for viability of the virus (Levinson *et al.*, 1982) but it is able to function in a manner which is largely independent of the location and orientation of the element within the viral genome (Fromm and Berg, 1983). Both enhancers are able to activate transcription in *cis* from homologous and heterologous promoters, relatively independent of position or orientation (Moreau *et al.*, 1981; Banerji *et al.*, 1981; deVilliers and Schaffner, 1981) and also to enhance transformation of cells in culture by other genes (Luciw *et al.*, 1983; Berg and Anderson, 1984). Transcriptional control elements having similar properties have been found in association with a number of viral and cellular genes (reviewed in Gluzman and Shenk, 1983). The function of the enhancer in early viral gene expression appears to be to increase the number of transcribing RNA polymerase II (POLII) molecules using the linked promoter (Treisman and Maniatis, 1985). A likely explanation of this finding is that enhancers are nucleosome-free regions which specifically bind factors involved in recruiting POLII, which after entering the chromatin, scans in both directions for a promoter (Moreau *et al.*, 1981), although alternative explanations involving alteration of template topology or binding to the nuclear matrix cannot be ruled out. Although the SV40 enhancer is capable of stimulating transcription about 10- to 15-fold *in vitro*, this is still an order of magnitude below its *in vivo* effect (Sassone-Corsi *et al.*, 1984; Sergeant *et al.*, 1984; Scholer and Gruss, 1985). In addition, *cis* activation is only seen *in vitro* if the enhancer element is positioned in close proximity to the promoter (Sassone-Corsi *et al.*, 1984; Sergeant *et al.*, 1984; Scholer and Gruss, 1985). Taken together, these results imply that higher order chromatin structure may render enhancers able to interact with promoters over large distances, possibly by enabling intervening DNA to be looped out, as proposed by Robbins *et al.* (1986). For Py, the sequences which activate DNA replication have been shown to lie within the region which defines the transcriptional enhancer (de Villiers *et al.*, 1984). The involvement of the Py enhancer in DNA replication is discussed in a later section.

Figure 1.2.

Functional organisation of the Py and SV40 central control regions. Components shown are implicated in early and late transcriptional control, and in the initiation of DNA replication by the results of genetic and DNA-protein binding studies described in the text. Nucleotide numbering is from left to right for Py, but from right to left for SV40. Enhancer domains A and B of Py (Herbomel *et al.*, 1984) and A-C of SV40 (Zenke *et al.*, 1986) are indicated. The proposed functions of these, of T-Ag binding sites A-C (Py) and I-III (SV40) and the Sp1 binding sites contained within the 21 bp repeats (SV40) are discussed in the text. Predominant early and late region cap sites are indicated with horizontal arrows. For each early region, EE indicates transcriptional start sites used early in lytic infection, and LE indicates those 5' ends found at later times.


Py
5050
5098
5131
5269
5297/0
27
76
86
116
121
164
175
A
B
ori
A
B
C
TATA
ATG
5077
5101
ENHANCER
5220
5250
148
153
LE
EE
LATE
EARLY
SV40
325
ENHANCER
LE
EE
39
31
5238
5231
C
72
72
III
TATA
II
I
ATG
251
179
107
21
21
22
B
A
A
350
0/5243
5163
ori — origin of DNA replication
— T-Ag binding site
100 bp

SV40 and Py enhancers both display some degree of host cell specificity. Thus the SV40 element has a higher activity in primate cells than in murine, rat, bovine or human cells (de Villiers *et al.*, 1982; Byrne *et al.*, 1983; Berg *et al.*, 1983; Spandidos and Wilkie, 1983; Augereau and Wasylyk, 1984). Conversely, while the wild type Py enhancer is active in differentiated mouse cells, it is much less so in undifferentiated mouse embryonal carcinoma (EC) cells (Herbomel *et al.*, 1984). Furthermore, Py variants which are able to grow in EC cells (reviewed in Levine, 1983) have mutations, duplications and/or deletions in the enhancer region. It seems likely that differences in enhancer activity observed in different cell types result from different activities of *trans*-acting factor(s) which interact with the enhancer. Further evidence for enhancer-specific factor(s) comes from studies using systems in which enhancer sequences can stimulate transcription both *in vivo* and *in vitro*. Such stimulation can be abolished by addition of competitor DNA containing enhancer sequences, suggesting that the enhancer interacts with factor(s) present in the transcription system (Scholer and Gruss, 1984; Sassone-Corsi *et al.*, 1985). Other viral and cellular enhancers are able to compete with the SV40 enhancer, whereas mutated enhancers which have lost their ability to stimulate transcription *in vivo* are also unable to compete efficiently for *trans*-acting factors *in vitro* (Wildeman *et al.*, 1984). Direct assays of DNA-factor binding have recently enabled the isolation of factors which bind to the Py (Piette *et al.*, 1985) and SV40 (Lee *et al.*, 1987) enhancers. One of the latter factors, termed AP1, also binds to the human metallothionein IIa basal level enhancer. The binding sites for these factors have been accurately mapped using DNA footprinting techniques. Such studies represent important steps towards the clarification of enhancer function.

The cell type specificity of enhancer action initially suggested that enhancers might be an important determinant of viral host range. Indeed, replacement of the Py enhancer with that of SV40 caused a 4- to 6-fold increase in T-Ag expression in primate cells (de Villiers *et al.*, 1982). However, in a separate study, a recombinant Py virus containing the SV40

enhancer region displayed DNA replication and gene expression in mouse cells, but not in primate cells (Campbell and Villarreal, 1985). Conversely, a hybrid SV40 virus containing the Py enhancer grew in monkey cells but not in mouse cells (Weber *et al.*, 1984). It would seem, then, that while the enhancer is able to alter the cell-type preference of a virus within a given species (see below), the enhancer alone is not able to alter the host range of a second virus.

Enhancer structure.

A recent study has defined the sequences responsible for the activity of the SV40 enhancer by using systematic deletions and point mutations (Zenke *et al.*, 1986). This analysis localised the minimum enhancer sequence to between nts 185 and 275, which includes one copy of a 72 bp repeat (Fig. 1.2), and showed that this region contained several short regions of homology with other enhancers which were important for activity. These include the enhancer "core" sequence TGGAAAGT and 8 bp segments of alternating purines and pyrimidines, capable of assuming the Z-DNA conformation. Within the SV40 enhancer these motifs are arranged into two domains, A and B, which separately show weak enhancer activity, but which, if combined, synergistically stimulate transcription (Zenke *et al.*, 1986). This synergism is maintained even if the domains are separated by over 50 bp or inverted with respect to one another. A third domain, C, is located further upstream from the early promoter, but its activity is masked in the presence of domains A and B. Similarly, the Py enhancer can be divided into two conceptual domains (Herbomel *et al.*, 1984), one (A) centered on a block of homology with the adenovirus E1a enhancer (AGGAAGTGACT), and the other (B) on a block of homology with the SV40 enhancer core sequence (TGTGGAAAG). Again, these motifs alone are not sufficient for enhancer function, but require additional flanking sequences. Comparison of many Py variants which can grow both in EC cells and in differentiated cells has shown that the sequences which constitute enhancer A are retained or duplicated

in all of them, whereas enhancer B has deletions and/or duplications (Melin *et al.*, 1985). One variant (PyF441) selected to grow in EC cells has only a single point mutation, in enhancer B (A to G at nt. 5258) (Fujimura *et al.*,1981). Together, these results suggest that a modified enhancer B is active in EC cells, and that this modification is sufficient to alter the viral host cell type range in this respect. A recent study investigated the fine structure of the Py enhancer by generating viruses with mutated, inactive enhancers and subsequently isolating revertants (Tang *et al.*, 1987). Reversion occurred by one of three mechanisms: by restoration of one of the previously mutated G residues between nts 5134 and 5140, within the adenovirus E1a homology; by acquisition of the A to G mutation at nt. 5258 described by Fujimura *et al.* (1981) or by duplication of mutated sequences between nts 5146 and 5292.

Interestingly, duplication of SV40 sequences that do not include either the enhancer core sequence or the alternating purine-pyrimidine blocks can restore enhancer function as well as viability to viral mutants lacking the entire enhancer region (Weber *et al.*, 1984; Swimmer and Shenk, 1984). It is difficult to reconcile these results with those of Zenke *et al.* (1986), but at this stage the consensus view is that both SV40 and Py enhancers consist of multiple elements, and deletion of one element can be compensated for by duplication of one or more of the remaining elements (Herr and Gluzman, 1985).

5' ends of SV40 and Py early mRNAs.

The positions of SV40 and Py early transcription initiation have been mapped by S1 nuclease and primer extension methods, and have been confirmed in certain cases by direct sequence determination of the capped mRNAs (Haegeman and Fiers, 1980; Kahana *et al.*, 1981; Ghosh and Lebowitz, 1981; Thompson *et al.*, 1979; Benoist and Chambon, 1981; Fromm and Berg, 1982; Heiser and Eckhart, 1982; Kamen *et al.*, 1982; Cowie *et al.*, 1982). For both viruses the capped early mRNAs have heterogeneous 5' ends, which probably represent heterogeneous transcription start sites (Gidoni *et al.*, 1981). These

sites map between nts 5231 and 5237 for SV40 and between nts 148 and 153 for Py (Fig. 1.2). Minor Py early 5' ends map to nts 14, 22 and 302. In addition, both viruses generate early strand RNAs with different 5' ends at late times in infection, after the onset of viral DNA replication. For SV40, these 5' ends map around nts 28 to 34, and for Py to around nts 5220 to 5250. Detailed analysis by Kamen *et al.* (1982) suggested that, for Py, these upstream RNAs are generated by endonucleolytic cleavage of long RNAs initiated at the major cap sites; this hypothesis was supported by the finding that these RNAs are not capped at their 5' ends. Overproduction of early strand transcripts after the onset of DNA replication of a ts-A mutant virus (see below) led to an accumulation in the nucleus of infected cells of early RNA, most of which had 5' termini at the upstream positions. It seems probable that late in infection termination of transcription from early promoters is inefficient, as is transcription from the late Py promoter. This inefficiency permits transcription to extend around the circular genome, generating giant RNAs. Subsequent cleavage of these primary transcripts could give rise to RNAs with upstream 5' ends without recourse to an alternative set of transcriptional start sites. In this light an unexpected finding was that inactivation of the downstream (EE) initiation sites led to the increased use of the upstream (LE) sites (Baty *et al.*, 1984). This is clearly inconsistent with the extension of the transcriptional readthrough hypothesis to SV40. Whilst it is known that the LE RNAs of Py are not capped, the analogous experiments have not been done for SV40. The LE RNAs are spliced and polyadenylated in both cases, but seem to be translated very poorly *in vitro* into T-Ags in the case of SV40 (Buchman *et al.*, 1984). This could reflect the absence of a cap structure, which may be involved in ribosome binding, or the presence of short reading frames upstream of the T-Ag coding sequences. These reading frames have the potential to encode 23 (SV40) and 53 (Py) amino-acid polypeptides, and could exclude the use of the downstream T-Ag AUGs (Kozak, 1983), but there is no evidence for the production of these polypeptides *in vivo*. The fact that, for SV40, production of the LE RNAs is absolutely correlated with the

presence of an active origin of replication, and not with the presence of large T-Ag, raises the intriguing possibility that LE RNAs are involved in activation of the origin, although the cause-effect relationship could also be reversed.

The early promoter.

In common with a large number of POLII genes whose expression is regulated, rather than constitutive, the early regions of SV40 and Py have a TATA box sequence (Breathnach and Chambon, 1981) 25 to 30 nts upstream from the predominant early transcriptional start sites. For SV40, site-directed mutagenesis has been used to change the TATA box sequence from TATTTAT to TATCGAT. In a transient assay, the mutation led to a 20- to 30-fold reduction in the amount of RNA synthesised from the EE start sites, but did not affect the levels of LE transcripts (Wasylyk *et al.*, 1983). While this was originally taken to be evidence for the existence of separate promoters for the LE and EE start sites, an alternative explanation is that the mutation causes heterogeneous transcription initiation at new sites other than EE, but that these transcripts can still be processed by the putative endonucleolytic mechanism which generates the LE 5' ends. In support of this conclusion is the finding that RNA isolated from cells transformed by a deletion mutant lacking the TATA box had more heterogeneous 5' ends than that from cells transformed by wild-type SV40 (Benoist and Chambon, 1981). In complementary experiments, deletion of the EE cap site region led to the generation of new transcriptional start sites about 30bp downstream. For SV40, then, the function of the TATA box seems to be to direct the initiation of transcription to about 30 bp downstream, and to mediate the high efficiency of initiation *in vivo*. In contrast, the Py TATA box does not seem to influence cap site position (Dailey and Basilico, 1985), and deletion of both the TATA box and the major cap site sequences led to efficient transcriptional initiation at other sites which are only used at a low level by the wild-type virus (Bendig *et al.*, 1980; Kamen *et al.*, 1982; Jat *et al.*, 1982; Mueller *et al.*, 1984). Thus, whilst the normal RNA initiation

sites are not required for efficient transcription by either virus, the TATA box seems to be significant in this respect only in SV40.

In addition to the TATA box and enhancer sequences described above, one further sequence domain has been implicated in SV40 early transcription initiation. This consists of two 21 bp repeats and an adjacent, related 22 bp sequence, imprecisely referred to as the three 21 bp repeats (Fig. 1.2), which together contain six copies of the sequence CCGCCC. In contrast, the sequences which control Py early region transcriptional initiation are not well defined. The only upstream sequences required for efficient transformation by Py are those which constitute the enhancer element (Jat *et al.*, 1982; Mueller *et al.*, 1984). SV40 mutants lacking the 21 bp repeats transform cells less efficiently than wild-type and produce reduced levels of T-Ags (Benoist and Chambon, 1981; Fromm and Berg, 1982, 1983; Everett *et al.*, 1983). The inversion of the 21 bp repeats at their normal location did not significantly alter the amount of T-Ag synthesis, but the 21 bp repeats are classified as promoter, rather than enhancer elements, since their potentiation of early transcription is not demonstrable over large distances. The effect of the 21 bp repeats on early transcription has been analysed more rigorously by the use of point mutations (Baty *et al.*, 1984), which suggested that all six CCGCCC sequences are required for absolutely normal gene expression. Whereas the two CCGCCC sequences proximal to the TATA box were essential for EE transcription starts, the four remaining sequences stimulated transcription from both EE and LE start sites.

Fractionation of cell extracts has permitted the partial or complete purification of several POLII transcription factors, some of which bind tightly to the SV40 early promoter. These include a TATA box binding factor (Davidson *et al.*, 1983; Parker and Topol, 1984) and a factor, Sp1, which is specifically required for transcription from the SV40 early promoter, but which is not needed for transcription from several other promoters, including that of the Py early region (Dynan and Tjian, 1983a). Sp1 has been shown to bind to the CCGCCC sequences in the 21 bp repeats by DNAseI footprinting

and methylase protection assays (Dynan and Tjian, 1983b; Jones and Tjian, 1984; Gidoni *et al.*, 1984). Comparison of Sp1 binding sites from a number of genes has led to the extension of the consensus binding sequence to (A/G)(C/T)(C/T)CCGCCC(A/C) (Kadonaga *et al.*, 1986). A good correlation between the DNA binding and transcription stimulatory activities indicated that Sp1 activates transcription from the SV40 early promoter by binding to the 21 bp repeats. A recent study suggests that it is important that the proteins which bind to the enhancer, 21 bp repeats and TATA box all lie on the same face of the DNA helix (Takahashi *et al.*, 1986). Thus alterations which shift the stereospecific alignments of these binding sites by half a turn of the helix have a much more drastic effect on transcriptional efficiency than those which alter the spacing by an integral number of turns.

Regulation of early transcription by large T-Ag.

The early promoters of both SV40 and Py are autoregulated by their respective large T-Ags. For both viruses, this was first suggested by the observation that early viral RNA was overproduced at the non-permissive temperature following infection with viruses carrying a temperature-sensitive (ts) large T-Ag allele (Tegtmeyer *et al.*, 1975; Reed *et al.*, 1976; Khoury and May, 1977; Cogen, 1978; Fenton and Basilico, 1982). The mechanism of the autoregulation phenomenon has been investigated by extensive *in vitro* studies (Tjian, 1981; Tenen *et al.*, 1983; Gaudray *et al.*, 1981; Pomerantz *et al.*, 1983; Pomerantz and Hassell, 1984; Cowie and Kamen, 1984; Dilworth *et al.*, 1984). These studies have identified three large T-Ag binding sites within the central control region of each virus (Fig 1.2). Transcription from viral mutants having an otherwise intact promoter but lacking T-Ag binding sites I and II (SV40) or A and B (Py) is not inhibited by large T-Ag (Rio and Tjian, 1983; Dailey and Basilico, 1985). These studies demonstrate that regulation of early transcription occurs through the interaction of large T-Ag with the viral DNA. Large T-Ag-mediated autoregulation does not seem to involve blockage or slowing of POLII progression along the template; when SV40 T-Ag binding sites are placed

downstream of a heterologous promoter, *in vitro* transcription is unaffected by the addition of exogenous large T-Ag, under conditions in which transcription from the SV40 early promoter was inhibited (Myers *et al.*, 1981). These results suggested that rather than interrupt the progression of POLII along the template, large T-Ag binding to its sites prevents binding of POLII to the template DNA. For the hybrid gene, once transcription had been initiated, it was able to proceed through the downstream sites to which large T-Ag had bound.

Splicing of early transcripts.

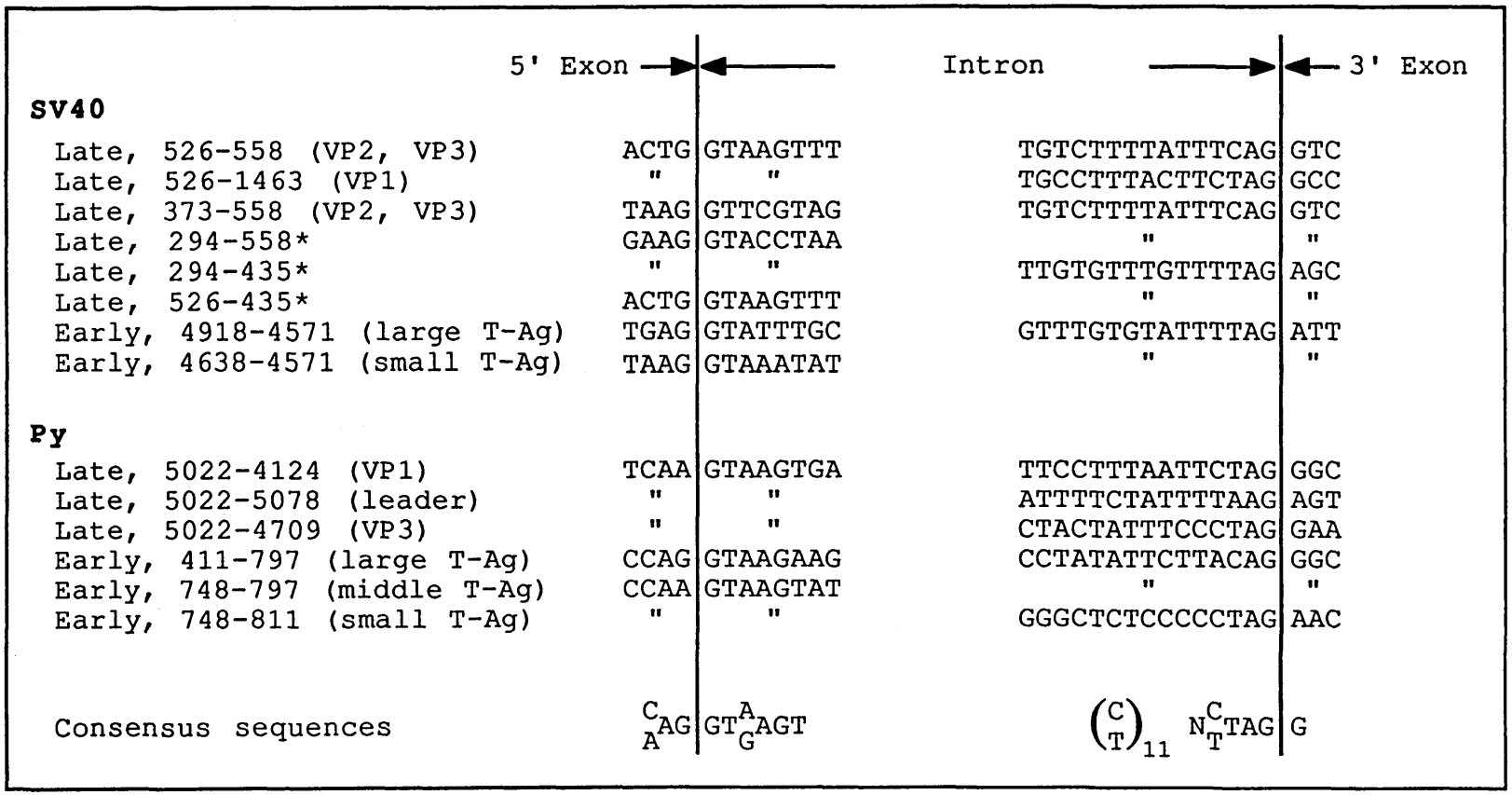
The amount of early mRNA that accumulates by 10 to 15 hours after Py or SV40 infection of permissive cells represents between 10^{-6} and 10^{-5} of the total cytoplasmic RNA (Sambrook *et al.*, 1973; Kamen *et al.*, 1975; Bacheler, 1977; Cogen, 1978). This is equivalent to between 10 and 100 early mRNA molecules per cell, compared with perhaps 100,000 to 400,000 molecules of cellular mRNA. Despite the low abundance of early mRNAs, it has been possible to characterise their structure at the nucleotide level by S1 nuclease and cDNA cloning analyses (Reddy *et al.*, 1979; Treisman *et al.*, 1981). For SV40, the mRNA coding for small T-Ag results from splicing the primary transcript between nts 4638 (5') and 4571 (3'). The mRNA encoding the large T-Ag uses the same 3' splice site, but uses an alternative 5' site at nt. 4918. In addition, a minor 5' splice site at nt. 4225 and multiple 3' sites around nt. 4100 have been implicated in the production of large T-Ags of reduced size (Sompayrac and Danna, 1985). Similarly, the three major Py early mRNAs share common sequences towards their 5' and 3' ends, but differ in the location and extent of the sequences which are spliced out. The large, middle and small T-Ag mRNAs are spliced between nts 411 and 797, 748 and 797 and 748 and 811 respectively (Fig 1.1). This differential splicing causes each of the processed mRNA products to be translated in a different reading frame after the position of the splice junction, and is crucial to the viral strategy of using overlapping reading frames.

The mechanism of splicing.

The intron/exon boundaries of SV40 and Py mRNAs correlate well with the consensus sequences derived by Mount (1982) (Table 1.2), and it is reasonable to suggest that the biochemistry of pre-mRNA splicing in both viruses is the same as that used for cellular transcripts. The recent development of *in vitro* splicing systems has permitted the elucidation of the mechanism of splicing in considerable detail (reviewed in Green, 1986 and Padgett *et al.*, 1986). These studies have shown that splicing proceeds in two stages. Firstly, cleavage at the 5' splice site occurs, and the 5' phosphorylated terminus of the intron is joined to the 2'-OH of an A residue located at the "branch site", between 20 and 40 nts upstream from the 3' splice site. It seems probable that these events occur in a concerted reaction. Genetic analysis of this branch site has identified, for yeast genes, a highly conserved consensus sequence (UACUAAC) located between 20 and 60 nts upstream of the 3' splice site (Langford *et al.*, 1984; Pikielny *et al.*, 1983). The A residue nearest the 3' end of this sequence is the branch point of the splicing intermediate, and point mutations at this position can inactivate splicing (Langford *et al.*, 1984). A survey of known intron sequences from higher eukaryotes has shown that there is only a low degree of homology with the yeast branch site consensus, though the sequence CU(A/C)A(U/C) is often found in the appropriate location (Keller and Noon, 1984). In contrast with the results obtained with yeast, deletion of the loosely conserved sequence from mammalian genes does not appear to prevent splicing *in vivo*, and the branch site still remains ill-defined. This first step generates a linear RNA corresponding to the upstream segment of the precursor and a looped, or "lariat" intermediate, a structure comprising the intron and the remaining downstream part of the pre-RNA. The two intermediates are presumably retained within a specific nuclear microenvironment, the "spliceosome", although there is evidence that splicing between exons located on separate pre-RNAs (*trans*-splicing) is possible (Konarska *et al.*, 1985; Solnick, 1985; Krause and Hirsh, 1987). Cleavage then takes place at the 3' splice site, and the 3' end of the

Table 1.2.

Sequences of 5' and 3' splice regions of SV40 and Py. Each known splice is described by the 5' and 3' nucleotide numbers at the junction, and by the protein product encoded, where appropriate. The vertical lines divide sequences included in the mature mRNAs from those which are spliced out. The sequences are aligned with the consensus derived by Mount (1982). Splices detected only at a low level in the lytic cycle of SV40 are denoted by asterisks. The sequences are taken from Reddy *et al.* (1978, 1979) (SV40) and from Treisman (1980) and Treisman *et al.* (1981b) (Py).


5' Exon
Intron
3' Exon
SV40
Late, 526-558 (VP2, VP3) ACTG GTAAGTTT TGTCTTTTATTTCAG GTC
Late, 526-1463 (VP1) " " TGCCTTTACTTCTAG GCC
Late, 373-558 (VP2, VP3) TAAG GTTCGTAG TGTCTTTTATTTCAG GTC
Late, 294-558* GAAG GTACCTAA " "
Late, 294-435* " " TTGTGTTTGTTTTAG AGC
Late, 526-435* ACTG GTAAGTTT " "
Early, 4918-4571 (large T-Ag) TGAG GTATTTGC GTTTGTGTATTTTAG ATT
Early, 4638-4571 (small T-Ag) TAAG GTAAATAT " "
Py
Late, 5022-4124 (VP1) TCAA GTAAGTGA TTCCTTTAATTCTAG GGC
Late, 5022-5078 (leader) " " ATTTTCTATTTTAAG AGT
Late, 5022-4709 (VP3) " " CTACTATTTCCCTAG GAA
Early, 411-797 (large T-Ag) CCAG GTAAGAAG CCTATATTCTTACAG GGC
Early, 748-797 (middle T-Ag) CCAA GTAAGTAT " "
Early, 748-811 (small T-Ag) " " GGGCTCTCCCCCTAG AAC
Consensus sequences $^{C}_{A}AG\ GT^{A}_{G}AGT$ $\left(^{C}_{T}\right)_{11}\ N^{C}_{T}TAG\ G$

upstream exon is joined to the 5' end of the downstream exon, again presumably in a concerted reaction. The intron is released as a lariat which, though relatively stable in the *in vitro* system, is rapidly degraded *in vivo*. The 5' splice site consensus sequence shows considerable complementarity to the 5'-terminal region of the small nuclear (sn) RNA U1, which has been shown to play an essential role in the splicing process (Padgett *et al.*, 1986). Further evidence suggests the involvement of U2, U4, U5 and U6 snRNPs in the formation, prior to splicing, of nucleoprotein complexes at the 5' splice region, the branch region and the 3' splice region.

Lewin (1980) suggested that the splicing machinery attaches at the 5' splice region and travels along the RNA until the first 3' splice site is encountered. This hypothesis was tested by constructing genes in which either the 5' or the 3' splice regions were duplicated, but the results were equivocal. In some cases, the distal copies of the duplicated splice sites were chosen (Kuhne *et al.*, 1983) whilst in similar experiments the proximal copies were selected (Lang and Spritz, 1983). It was later shown that the length and sequence of the exons adjoining the sites play a role in splice site selection, perhaps because of their effect on higher-order RNA structure (Reed and Maniatis, 1986). In addition, *trans*-acting factors may modulate the availability of splice regions for splicing. In this light it is intriguing that, of the four conceivable combinations of 5' and 3' splice sites in the Py early region, one combination, that of the 5' site at nt. 411 with the 3' site at 811, does not appear to be used to any extent (Treisman *et al.*, 1981). Similarly, the late mRNAs of SV40 use two predominant donors and two acceptors, and again one conceivable combination of these sites, that of the donor at nt. 373 with the acceptor at nt. 1463, does not seem to be used (Table 1.2).

Differential Splicing of Cellular Transcripts.

The potential of one transcription unit to encode more than one product by differential splicing of the available exons, first demonstrated for SV40 and Py, has since been shown

to be exploited by cellular transcription units. Variations in the protein products of a transcription unit can be brought about where, as in Py and SV40, alternative splice donor or acceptor sites are used for a particular exon. This phenomenon has been described for a number of cellular genes, including those encoding chicken ovomucoid (Stein *et al.*, 1980), human granulocyte colony-stimulating factor (Nagata *et al.*, 1986), human growth hormone (De Noto *et al.*, 1981), rat prolactin (Maurer *et al.*, 1981) and rat fibronectin (Tamkun *et al.*, 1984). In addition to this relatively subtle mechanism, "exon selection" may be used to generate greater diversity of products from a particular transcription unit. For example, the membrane-bound and secreted forms of IgM are encoded by different combinations of exons driven by a single promoter (Alt *et al.*, 1980; Early *et al.*, 1980). In early-stage B cells both secreted and membrane-bound μmRNAs are produced, while secreted μmRNA is the major species in plasma cells. The two mRNAs have five common 5' exons, followed by, for the secreted μmRNA, a sixth exon ending in a poly(A) site. The membrane-bound μmRNA splices out the latter site, and consequently has a shorter sixth exon; this is followed by two further exons, ending in a different poly(A) site. An essentially similar arrangement is found in the rat calcitonin gene, which has six exons available for differential splicing to generate two mRNA species (Amara *et al.*, 1982). In thyroid C cells, the mRNA produced encodes the precursor of calcitonin, and contains the first four exons, the fourth ending in a poly(A) site. In contrast, in brain most or all of the mRNA contains the first three exons spliced to exons 5 and 6, and encodes a calcitonin gene related peptide (CGRP), which appears to differ in function from calcitonin. As in the IgM transcription unit, the alternative splicing pathway removes an alternative poly(A) site contained within an intron from the mature mRNA. Analogous results have been obtained for the Drosophila GAR transformylase gene, which may produce two proteins, derived by alternative RNA processing, to channel metabolic intermediates into two different pathways (Henikoff *et al.*, 1983), and for the transcription unit encoding nonmuscle and muscle tropomyosin isoforms (Helfman *et al.*, 1986). In these systems,

where there is differential splicing of alternative 3' exons, differential polyadenylation and splicing are inextricably linked. Examples of differential splicing not involving 3' exons include the mouse myelin basic protein gene, which consists of seven exons and encodes four related protein products by the inclusion or exclusion of exons 2 and/or 6 (De Ferra *et al.*, 1985). An extreme case is that of the rat troponin T gene, which has 18 available exons and encodes at least ten protein isoforms through differential splicing. Processed mRNAs display the mutually exclusive inclusion of either exon 16 or exon 17, as well as the inclusion of all conceivable combinations of exons 4 to 8 (Breitbart *et al.*, 1985). Furthermore, the patterns of alternative splicing differ depending on cell type, with *trans*-acting factors required for correct alternative splicing being induced during myogenesis (Breitbart and Nadal-Ginard, 1987).

Early mRNA polyadenylation.

In higher eukaryotes, the 3' end of a processed mRNA is derived by endonucleolytic cleavage of a longer precursor molecule, and the majority of mRNAs are further modified by the addition of a poly(adenylic acid) (poly(A)) tract, between 50 and 200 nucleotides in length, to the exposed 3' terminus. All three predominant Py early mRNAs are polyadenylated after the 3' end of the large T-Ag coding sequences, in the vicinity of nt. 2940 (Kamen *et al.*, 1982; Heiser and Eckhart, 1982; this study). In addition, a minor set of RNAs polyadenylated in the middle of the large T-Ag sequences, in the vicinity of nt. 1525 has been observed in lytically infected and some transformed cells (Treisman *et al.*, 1981; Fenton and Basilico, 1981; Heiser and Eckhart, 1982; this study). In contrast, all SV40 early mRNAs have a common polyadenylation site 3' to the large T-Ag translational termination codon, around nt. 2964 (Fig. 1.1).

Endonucleolytic cleavage and poly(A) addition.

Evidence that polyadenylation involves two steps, cleavage and poly(A) addition, first

came from studies on the adenovirus type 2 (Ad2) late mRNAs, which are cleaved and polyadenylated soon after the RNA polymerase passes the poly(A) site and before the primary transcript is completed (Nevins and Darnell, 1978). It was reasoned that, since a free 3' end would not be available for attack by a putative exonuclease before the primary transcript was complete, the transcript must be cleaved endonucleolytically prior to poly(A) addition. This interpretation has since been substantiated by the results of *in vitro* studies with the Ad2 L3 transcript (Moore and Sharp, 1985). In this system, in the absence of Mg^{2+}, it has been possible to isolate the downstream RNA segment which results from precise endonucleolytic cleavage. The cleavage reaction is rapidly followed by the polymerisation of an initial poly(A) tract of about 130 nts, which is extended to about 250 nts on prolonged incubation. It appears that cleavage is the rate-limiting step in cleavage/ polyadenylation. Thus an Ad2 E1a gene with a mutated poly(A) site gives only a small number of accurately cleaved 3' ends when transfected into human HeLa cells, but all of these ends are polyadenylated (Montell *et al.*, 1983). While it would seem that this result suggests that cleavage is the rate-limiting step, an alternative interpretation is that non-polyadenylated 3' ends are very unstable (Birnstiel *et al.*, 1985). Furthermore, while the cleavage and poly(A) addition steps are closely associated, there is no obligatory mechanistic link between the two. This is clear because, by using the analogue inhibitor cordecypin triphosphate (3'dATP), Moore and Sharp (1985) were able to demonstrate cleavage *in vitro* in the absence of poly(A) synthesis. The cell extract used by Moore and Sharp differed from extracts previously used for *in vitro* polyadenylation (Manley, 1983; Chen-Kiang *et al.*, 1982) in that it was capable of processing exogenously added transcripts, suggesting that POLII transcription and polyadenylation need not necessarily be linked. That this is true has been demonstrated *in vivo* using a recombinant gene containing an RNA polymerase III (POLIII) promoter, the Ad VAI promoter, fused to the body and 3' flanking sequences of a gene normally transcribed by POLII, the herpesvirus tk gene (Lewis and Manley, 1986). Following transfection into human 293 cells, a high

proportion of the POLIII transcripts from this gene were accurately cleaved and polyadenylated.

Sequence requirements for polyadenylation: AAUAAA.

Since 3' cleavage appears to be the critical step in 3' processing, selection of the site of cleavage would be expected to be a highly ordered process. The identification of components of the cleavage/polyadenylation machinery and, in particular, the RNA sequences with which this machinery must interact has been an active area of research in recent years. The comparative study of a number of POLII genes has revealed the presence of a highly conserved AAUAAA consensus sequence, usually located 10 to 30 nts upstream of the site of cleavage/polyadenylation in the pre-mRNA (Proudfoot and Brownlee, 1976). The poly(A) tract very often begins immediately after a point where there is a CA or TA dinucleotide in the template, such that the first A residue in the tract could be transcribed, rather than added post-transcriptionally. While polyadenylation is almost always found at a single position downstream of the hexanucleotide consensus, three exceptions to this rule have been documented. In the first, oligonucleotide-primed sequence analysis of mRNA from and cDNA clones of the bovine prolactin gene was used to demonstrate that polyadenylation in this gene is heterogeneous, occurring at at least three positions: 17, 22 and 29 nts downstream from the AAUAAA sequence (Sasavage *et al.*, 1982). Secondly, S1 nuclease analysis of 3' ends and sequencing of processed pseudogenes showed that, for the mouse ribosomal protein L30 gene, the position of poly(A) addition varies between 14 and 28 nts downstream of the hexanucleotide (Wiedemann and Perry, 1984). The third case is that of the hepatitis B virus surface antigen gene, which has mRNA 3' termini between 12 and 19 nts downstream from the sequence UAUAAA (Simonsen and Levinson, 1983). The reasons for the heterogeneity of polyadenylation in these systems is not known, but may relate to the components other than the AAUAAA sequence which contribute to the generation of a functional poly(A)

site and which are alluded to below.

That the ubiquity of AAUAAA reflected a functional role for this sequence in cleavage/polyadenylation was first established for the SV40 late transcription unit (Fitzgerald and Shenk, 1981). A 16 bp deletion of sequences including the hexanucleotide consensus prevented late mRNA polyadenylation after infection of monkey CV-1 cells. A second class of mutants lacked segments between the AAUAAA and the normal poly(A) site. These mutants produced mRNAs polyadenylated at new sites, progressively further downstream on the wild-type sequence as the size of the deletion was increased. For these mutants, polyadenylation always occurred at a single site between 11 and 19 nts downstream from the AAUAAA sequence, with cleavage downstream from the dinucleotide CA clearly preferred. A double deletion mutant lacking sequences both upstream and downstream from the AAUAAA sequence produced a class of mRNAs polyadenylated at the CA immediately following the hexanucleotide, as well as other mRNAs polyadenylated at up to 16 nts downstream. This was the only mutant described whose pattern of polyadenylation loosely resembles that of the three heterogeneously polyadenylated genes described above. It is interesting to note that none of the mutations surrounding the AAUAAA sequence changed the steady-state levels of late mRNAs produced. Insertion of sequences derived from elsewhere in the SV40 genome into the region between the hexanucleotide and the wild-type poly(A) addition site caused polyadenylation to occur at novel CA and GA dinucleotides 15 and 20 nts downstream from the AAUAAA. Similar findings resulted from a study of the Ad E1a poly(A) site, in that mutation of the U residue in the AAUAAA sequence to a G caused a dramatic reduction in the level of RNA cleavage at the normal site (Montell *et al.*, 1983). More profoundly significant was the finding that mutation of the sequence AAUAAA to AAUAAG was solely responsible for the reduction in activity of a human α-2 globin gene in α-thalassaemia patients (Higgs *et al.*, 1983). When the cloned defective α-2 allele was transfected into human HeLa cells, the majority of its transcripts extended beyond the

mutant poly(A) site. Taken together, these results showed that the AAUAAA sequence is important for the efficient generation of mature mRNA 3' ends, and that it is also implicated in determining the site of cleavage.

A survey of published poly(A) sites shows that, while approximately 80% of mRNA sequences contain the AAUAAA sequence just upstream of their poly(A) tract, some variation in this consensus seems to be tolerated. Apparently functional variant hexanucleotides include: AUUAAA, AAUUAA, AAUACA, AAUAAU, AAUAAC, CAUAAA (Wickens and Stephenson, 1984), UAUAAA (Simonsen and Levinson, 1983), UAUAAG (Setzer *et al.*, 1980) and AGUAAA (Donehower *et al.*, 1981; Capon *et al.*, 1983).

The study of SV40 late AAUAAA mutants injected into *Xenopus* oocytes has shown that four different point mutations reduce cleavage efficiency. The reduction, calculated by comparing the fraction of mutant RNA that is processed to the fraction of wild-type RNA that is processed, is by a factor of 8 for AAUACA, by a factor of 25 for AAUUAA, and by a factor of 50 or more for AACAAA and AAUGAA (Wickens and Stephenson, 1984; inaccurately reported in the review by Birnstiel *et al.*, 1985). As in the other studies mentioned above, where cleavage efficiency was reduced by AAUAAA variants, all of the cleaved pre-RNA produced seemed to be polyadenylated. This is fairly compelling evidence for cleavage being the rate-limiting step, since, in the oocyte system, cleaved but unpolyadenylated RNAs are thought to be stable, and would be detected if present in any great quantity. It is interesting to note that the AAUACA and AAUUAA variants appear to function *in vivo*, at least at a low level, and together account for 4% of the known AAUAAA-like sequences near mRNA 3' ends, despite their apparent inactivity *in vitro* (Wickens and Stephenson, 1984). It seems possible that less efficient variants of AAUAAA could be used in some transcription units as a means of controlling gene expression, but in the absence of direct proof this possibility remains purely speculative.

The fundamental importance of the AAUAAA motif strongly suggested that component(s) of the cleavage/polyadenylation machinery might interact specifically with

the hexanucleotide. This has been confirmed by studies *in vitro* using cell extracts and synthetic pre-RNAs. A HeLa cell nuclear extract prevented hybridisation of an antisense oligodeoxynucleotide probe to the AAUAAA sequence of the late SV40 poly(A) site, as judged by the abolition of sensitivity to RNAse H (Ruskin and Green, 1985; Zarkower and Wickens, 1987). Protection in this assay was sequence-specific, being prevented by a point mutation giving AAUACA. Complex formation on synthetic RNAs that ended in the poly(A) addition site ("pre-cleaved" RNAs) also required the AAUAAA sequence, as did *in vitro* polyadenylation of such substrates (Manley, 1983; Zarkower *et al.*, 1986). It seems that the same component(s) recognised AAUAAA in cleavage and polyadenylation, since pre-cleaved RNA competed successfully in *in vitro* cleavage reactions. In contrast, polyadenylated RNAs did not form a complex, and competed less effectively for the AAUAAA-specific factor(s) than did pre-RNAs.

When polyadenylation reactions were incubated with RNAse T1 and antibodies directed against components of small nuclear riboproteins (snRNPs), fragments of RNA containing the AAUAAA and adjacent sequences (from Ad2 L3 or E2a and from SV40 early and late poly(A) sites) were again protected from digestion, and in this case were also immunoprecipitated (Hashimoto and Steitz, 1986). These results suggested that the factor that associates with AAUAAA might be a Sm snRNP containing a U RNA, and this conclusion was further supported by the finding that incubation with micrococcal nuclease, which extensively degraded the abundant small RNAs in the nuclear extract, completely inhibited polyadenylation of the L3 pre-RNA. An involvement of snRNPs in polyadenylation had earlier been suggested by the complementarity of sequence between U4 RNA and AAUAAA (5/6 match) and downstream sequences near certain poly(A) sites (Berget, 1984), and by the finding that accurate polyadenylation *in vitro* was inhibited by anti-Sm antibodies as well as antisera specific for U1 snRNPs (Moore and Sharp, 1984, 1985). In an analogous situation, the sea urchin snRNP containing the U7 RNA is required for formation of the 3' end of sea urchin histone H3 mRNA (Strub *et al.*, 1984).

Although the majority of histone mRNAs are not polyadenylated, their 3' ends appear to be generated by endonucleolytic cleavage (Birchmeier *et al.*, 1984; Krieg and Melton, 1984), as are 3' ends which become substrates for the poly(A) polymerase. Since the U7 sequence is complementary, in part, to the conserved, essential sequences adjacent to the site of histone pre-RNA cleavage, it has been suggested that the two RNAs interact in some way during the processing reaction (Strub *et al.*, 1984).

Other sequences involved in poly(A) site formation.

Although AAUAAA is the only rigidly conserved sequence which has been found adjacent to poly(A) sites, it is clear that this sequence alone is insufficient to elicit cleavage. This is evidenced by the presence of the hexanucleotide within the mRNA bodies of a number of transcription units, including the SV40 early region (Table 1.3). The fact that such internal AAUAAA sequences are not functional in polyadenylation leads one to ask what are the additional structural features of the pre-RNA which are involved at true poly(A) sites. This was the question posed by Simonsen and Levinson (1983), who constructed a series of deletion mutants of the hepatitis B virus surface antigen gene and transfected them into monkey COS cells. The results indicated that the AAUAAA-related hexanucleotide UAUAAA was able to direct efficient processing only when an additional 30 nts downstream from the signal were present. Similarly, although the deletion analysis of Fitzgerald and Shenk (1981) had shown that removal of sequences between 2 and 16 nts downstream from the SV40 late AAUAAA sequence had no major effect on polyadenylation efficiency, a more extensive deletion, removing sequences between 3 and 60 nts downstream, led to a dramatic decrease in efficiency (Sadofsky and Alwine, 1984). The phenotype of this mutant was an accumulation of extended late transcripts, which appeared to have an increased half-life due to the less efficient cleavage at the normal poly(A) site. Analogous results were obtained for the rabbit β-globin gene (Gil and Proudfoot, 1984) and the Ad2 E2a gene (McDevitt *et al.*, 1984), which both required an

Table 1.3.

Sequence comparison of functional, partially functional and non-functional AATAAA environments. 60 nt. segments of the Py and SV40 genomic sequences are aligned by the AATAAA sequence which is common to all of them. Py' and SV40' indicate that the sequence is taken from the late-sense strand (Py) or the early-sense strand (SV40). The number in each case refers to the nt. position of the first A of the hexanucleotide. The first five sequences, when transcribed into RNA, function as poly(A) sites *in vivo* for the transcription units indicated on the right hand side. The next two sequences occur within the early mRNA coding sequences of SV40, but are not used as poly(A) sites. The five remaining sequences are not represented in mature mRNAs. Every occurrence of the trinucleotide TGT is over-lined. In addition, (T+G) tracts of 4 nts or longer are underlined.

Position				Identity
Py 1478	TGTCTCATGCTATTTATTCT	AATAAA	ACGTTCCCGGCATTTCTAGTATACTCCACCAAAG	Alternative early
Py 2920	GCTGAATACACAGTTTATTG	AATAAA	CATTAATTTCCAGGAAATACAGTCTTTGTTTTTC	Major early
Py'2914	CTGGAAATTAATGTTTATTC	AATAAA	CTGTGTATTCAGCTATATCACATATGTCCTTCAA	Late
SV40'2607	ATAGCATCACAAATTTCACA	AATAAA	GCATTTTTTTCACTGCATTCTAGTTGTGGTTTGT	Early
SV40 2657	TTGTAACCATTATAAGCTGC	AATAAA	CAAGTTAACAACAACAATTGCATTCATTTTATGT	Late
SV40'2636	TGCAGCTTATAATGGTTACA	AATAAA	GCAATAGCATCACAAATTTCACAAATAAAGCATT	(Within early mRNAs)
SV40'3274	ACTTAGAAAAGAAACACCTA	AATAAA	AGAACTCAAATATTTCCCCCTGGAATAGTCACCA	(" " ")
Py'1474	TGCCGGGAACGTTTTATTAG	AATAAA	TAGCATGAGACAAATACCCAGTAAGGCTGCTAGG	(Not within mature mRNAs)
SV40 18	GCGGCCTCGGCCTCTGCATA	AATAAA	AAAAATTAGTCAGCCATGGGGCGGAGAATGGGCG	(" " " ")
SV40'1391	CCAGGAGCCTCAAATTTTTC	AATAAA	TTCACCTGACTGCACATTAGGACTTTGGCTTTGA	(" " " ")
SV40'962	CAAAGGTTTGTACTCCAGGA	AATAAA	ATATCATAGTAATCATCTGGCCTATACAAATCTA	(" " " ")
SV40'553	AGCAGCACCCATGGACCTGA	AATAAA	AGACAAAAAGACTAAACTTACCAGTTAACTTTCT	(" " " ")

additional 35 nt. region downstream from the AAUAAA sequence for mRNA 3' end formation, for the Ad2 E3a gene, in which sequences between 7 and 72 nts of an AUUAAA sequence were required (Bhat and Wold, 1985), for the herpes simplex virus tk gene, which needed sequences located 20 to 38 nts 3' to the first of two AAUAAA hexanucleotides (Cole and Stacy, 1985), and finally for the SV40 early region, which required sequences between 5 and 18 nts downstream from the cleavage site for efficient polyadenylation both *in vivo* (Hart *et al.*, 1985a) and *in vitro* (Hart *et al.*, 1985b). In contrast, a mutant bovine growth hormone gene containing only one authentic nucleotide downstream from the AAUAAA gave rise to transcripts which were polyadenylated efficiently, albeit at a novel site (Woychik *et al.*, 1984).

Searches of known poly(A) site sequences have been used to supplement the deletion analyses described above in attempting to identify auxiliary sequences involved in polyadenylation. At a stage when only a relatively small number of poly(A) site sequences was available, it seemed that a sequence closely related to UUUUCACUGC was present just downstream from the AAUAAA sequence of most genes (Benoist *et al.*,1980). As more sequences became available, however, the generality of this conclusion was lost and new consensus sequences were proposed. Berget (1984) noted the relatively frequent occurrence of the sequence CAYUG (where Y is a pyrimidine residue) in positions either 5' or 3' to the site of cleavage, though it now seems that this sequence too is not common to all poly(A) sites. The frequent occurrence of UUUU or UUGU motifs beyond the cleavage site, first noted by Lai *et al.* (1979) was extended to a consensus YGUGUUYY by McLauchlan *et al.* (1985), but close examination of the data shows that this sequence is present in only four of the ninety-five genes analysed. Much more highly represented is the simplified consensus UGU, which is found between 20 and 40 nts 3' to the AAUAAA consensus in seventy-two of the ninety-five genes. The UGU sequence is often found in close association with short stretches of G and U residues (Cole and Stacy, 1985). It is interesting in the light of this finding to return to the functional and non-functional

AAUAAA environments in Py and SV40 (Table 1.3). Of the twelve such sequences examined, the UGU sequence is present within the 40 nts downstream from the AAUAAA sequence only in transcripts spanning the four major poly(A) sites. The alternative early Py poly(A) site, which is only used at a relatively low level in the viral productive cycle, does not contain a downstream UGU sequence, which is also absent in the regions downstream from fully non-functional AAUAAA sequences. It would certainly seem that, within these viruses, the correlation between the presence of a UGU motif between 24 and 34 nts downstream from the AAUAAA sequence and the functionality of the latter is absolute. Furthermore, SV40 late region deletions which retained poly(A) site efficiency also retained the UGU sequence (Fitzgerald and Shenk, 1981; Sadofsky and Alwine, 1984). Similarly, those sequences defined by deletion analysis as essential for efficient polyadenylation in the seven genes described above all contain the UGU trinucleotide between 16 and 41 nts downstream from the AAUAAA homology. In the case of the herpes simplex virus tk gene, two UGU sequences are embedded in a 19 nt. region which complements the AAUAAA function and which contains 18 U or G residues (Cole and Stacy, 1985).

Two recent studies have defined accessory sequences necessary for polyadenylation by reconstruction of templates using synthetic oligodeoxynucleotides together with fragments of wild-type poly(A) sites. Hart *et al.* (1985b) have shown that sequences important in addition to AAUAAA for efficient cleavage at the Ad2 E2a poly(A) site could restore function to a defective SV40 early site lacking analogous, UGU-containing sequences. In further experiments, synthetic sequences containing (U+G)-rich motifs from the E2a and SV40 early genes were able to restore efficient cleavage to the deleted SV40 site, whilst inversion of these sequences completely abolished function (McDevitt *et al.*, 1986). Each motif contained a single copy of the UGU trinucleotide. A series of base-substitution variants of the (U+G)-rich motifs was used to further define function. For the E2a sequence, the UGU trinucleotide proved to be dispensable, indeed a run of

eight U residues was even better at restoring function than the wild-type octamer UUGUUUUU. This contrasted with the results with the SV40 sequence(GUUGUGGU, Table 1.3), when removal of the one remaining UGU sequence by a G to U change effectively abolished activity. The view that emerges from these results, in the context of the mass of data described earlier, is that at least two distinct elements, the U-rich sequence like that of the E2a gene, and a UGU-rich sequence like that of the SV40 early region, are able to combine with AAUAAA in the generation of a functional poly(A) site. Further evidence supports this notion of two distinct elements. The SV40 element retained activity when displaced 32 nts downstream from its normal site, but the E2a element did not function in this position. The E2a element was, however, the more effective of the two when at a position 7 nts downstream from the normal SV40 element location. These differences suggest that different recognition events might be involved for each element. A similar reconstruction approach has been used to identify accessory elements downstream from the rabbit β-globin gene AAUAAA sequence (Gil and Proudfoot, 1987). Sequence homology with other poly(A) sites revealed three candidate elements within the 35 nt. region shown to be essential for cleavage/polyadenylation by deletion analysis (Gil and Proudfoot, 1984): a CAUUG sequence (Berget, 1984), a UGU-containing element, UGUGUUGGA, identified in a number of genes by Taya *et al.* (1982), and a U-rich sequence UUUUUUGUGU which, incidentally, also contains two copies of the UGU trinucleotide. The results in this case seemed to differ from those of McDevitt *et al.* (1986), in that both the U-rich element and the UGU-containing element were required in addition to AAUAAA to restore efficient processing, but it is not yet known if the UGU components of the U-rich sequence are dispensable in this case. The CAUUG element, on the other hand, was unable to complement the activity of either of the two other sequences. Again, the effect of the UGU-containing and U-rich elements was dependent on their position, with the level of 3' end formation decreasing with increasing distance between the two motifs. The non-essential nature of the CAYUG motif has also been

demonstrated for the Xenopus β-1 globin gene (Mason *et al.*, 1986), but in this case, as with the bovine growth hormone gene (Woychik *et al.*, 1984), no accessory sequences necessary for efficient cleavage were identified, although removal of sequences including a (U+G)-rich element downstream from the AAUAAA sequence altered the position of poly(A) addition.

Insofar as it is possible to generalise, one must conclude that, while the AAUAAA-related sequence is of prime importance in cleavage/polyadenylation, other accessory sequences including U-rich and UGU-containing elements are implicated in poly(A) site definition. The fact that some sites function efficiently in the absence of these motifs may simply reflect the diverse possibilities for gene regulation through the interaction of different modulatory factors with the polyadenylation machinery.

Py and SV40: termination of early transcription.

Examination of Py early nuclear RNA pulse-labelled with ^{3}H-uridine, by hybridisation to restriction fragments of Py bound to nitrocellulose filters showed that about 80% of transcripts were terminated near the end of the early region, with about 20% apparently reading through into late region sequences (Acheson and Mieville, 1978). Since all eukaryotic mRNA 3' ends appear to be generated by endonucleolytic cleavage of a longer precursor, transcription beyond the poly(A) site must be a general phenomenon. The evolution of endonucleolytic mechanisms for generating precise mRNA 3' termini has relieved the eukaryotic cell of the requirement to precisely terminate transcription itself, which is the means by which RNA 3' ends are generated in prokaryotes (reviewed by Platt, 1986). The fact that no discrete RNA 3' ends have been detected beyond the early poly(A) sites of Py and SV40 is consistent with an imprecise transcription termination mechanism, but an alternative explanation would be that the pre-RNAs, though precisely terminated, are rapidly cleaved at the poly(A) site, and the resulting uncapped RNA fragment is very unstable.

Termination of transcription in other systems.

The early transcription units of Py and SV40 are typical of a large number of eukaryotic genes in that transcription is equimolar from the cap site to beyond the mature 3' end. Nevertheless, premature termination of transcription, first reported in bacteria (reviewed by Yanofsky, 1981), now seems also to be a feature of some eukaryotic genes. Such "attenuation" is suggested by the appearance *in vitro* of short RNAs derived from the promoter-proximal portions of the murine minute virus (Ben-Asher and Aloni, 1984) and the SV40 late region (Hay and Aloni, 1984), and by the finding of prematurely terminated Ad2 late RNAs both *in vitro* and *in vivo* (Maderious and Chen-Kiang, 1984). Similarly, premature termination of transcription accounts at least in part for a dramatic decrease in c-*myc* mRNA levels during differentiation of human promyelocytic leukaemia HL60 cells (Bentley and Groudine, 1986; the possible involvement of controlled termination of transcription in gene regulation is discussed in a later section). Examination of the template sequences at eukaryotic attenuation sites has suggested that premature termination of transcription at these sites resembles the rho-independent termination events observed in bacteria, where pausing at RNA hairpins that are followed by multiple uridine residues results in dissociation of the polymerase from the template. It seems likely that precise mechanisms of this sort do not account for termination of POLII transcription downstream from the mature 3' ends of mRNAs, an event which is probably important in preventing readthrough to adjacent genes. For such intergenic transcription of POLII transcription there exists the possibility of termination events linked to, or dependent upon, prior pre-RNA processing. Whilst the 3' processing event with which one is primarily concerned is polyadenylation, valuable information concerning POLII termination has come from the study of genes which use alternative 3' processing strategies.

The 3' ends of U1 snRNAs are generated by removal of a few 3'-terminal nucleotides from a pre-RNA transcribed by POLII (Chandrasekharappa *et al.*, 1983). Formation of

the pre-URNA 3' end is directed by a short, conserved sequence $GTTTN_{0-3}AAAPuNNAGA$ found between 9 and 19 nts downstream from the mature 3' end (Hernandez, 1985), though it is not yet clear that this sequence constitutes a POLII termination signal. Interestingly, in this system RNAs with a U1 3' end are only detected when the conserved sequence is placed downstream from a compatible snRNA promoter and enhancer; promoters from other genes cannot substitute for the snRNA element (Hernandez and Weiner, 1986). This result suggests that recognition of a termination signal might depend upon prior assembly at the 5' end of a gene of a transcription complex containing components later involved in site-specific termination downstream. In this way different promoters could recruit different terminator-recognition factors, and controls exerted through a particular promoter could be modulated to influence the position of transcription termination within or distal to the linked gene.

In experiments using a sea urchin histone H2A gene, whose mature mRNA 3' end is formed by a processing event directed by a terminal inverted repeat sequence, preliminary deletion analysis identified a 130 nt. (A+T)-rich element which appeared to cause POLII termination (Birchmeier *et al.*, 1984): in mutant genes lacking this spacer transcriptional readthrough to a downstream H1 gene was observed. This system also provided evidence to suggest that transcriptional termination need not necessarily be linked to 3' endonucleolytic cleavage, since removal of the terminal inverted repeat did not cause readthrough (Birchmeier *et al.*, 1984). Furthermore, the terminator element was active in human cells, in which faithful 3' processing of the sea urchin gene is not observed (Johnson *et al.*, 1986). In contrast to the U1 snRNA 3' element, the H2A terminator appears to be active when placed downstream from a heterologous promoter (Johnson *et al.*, 1986; Proudfoot, 1986), and so probably does not depend upon inclusion of specific recognition elements during assembly of the transcription complex. A more detailed analysis of the sea urchin H2A terminator has shown that the functionally important signals are both sequence-defined and complex (Johnson *et al.*, 1986). Surprisingly,

deletion analysis revealed that three distinct sequence elements were necessary and sufficient for prevention of transcriptional readthrough to a poly(A) site placed downstream. The first of these elements is contained within the H2A coding region and displays no obviously unusual sequence features or homologies with other putative POLII terminators. The second element, contained within a 60 nt. segment of 90% (A+T) content, is located just downstream from the H2A 3' terminal hairpin and processing signals, and contains two overlapping copies of the AATAAA poly(A) signal consensus, though this may simply reflect the unusual local base composition. The third essential component of the H2A terminator is located some 250 nts further downstream, and is within a 120 nt. region containing several oligo(T) stretches, notably a T_8 sequence, as its only obviously unusual sequence features. In this system transcription appeared to terminate heterogeneously within a region of several hundred nucleotides, and in this respect the findings are similar to those derived from the study of a number of other POLII genes.

In general, pre-RNAs destined for polyadenylation are terminated at heterogeneous sites within a region between a few hundred nucleotides and a few kilobases downstream from the site of cleavage (reviewed by Birnstiel, 1985). This conclusion is drawn from a considerable number of studies in which pulse-labelled nascent transcripts were tested for their ability to hybridise to cloned DNA fragments representing successively more promoter-distal sequences. Where the level of transcription is high, such analyses have been performed using pulse-labelled whole cells, but for the majority of genes a semi-*in vitro* transcription run-on assay has been used with isolated nuclei. A progressive reduction in hybridisation levels of nascent transcripts to the progressively more promoter-distal DNA fragments is interpreted as evidence of heterogeneous transcriptional termination. Although differences in the rate of *in vitro* elongation might also contribute to differences in incorporation, the polar nature of the hybridisation fall-off supports an interpretation based on termination, rather than pausing. In four such cases the involvement of specific sequence elements in the termination function has been

investigated.

In the case of the chicken ovalbumin gene, polymerase loading was found to drop off abruptly within a 168 nt. region beginning 826 nts downstream from the predominant site of cleavage/polyadenylation (LeMeur *et al.*, 1984). The 168 nt. fragment is not obviously unusual in its sequence, which is 55% (A+T)-rich, and the competence of this sequence in directing POLII termination, either alone or in association with other elements, has yet to be reported.

When 3' sequences from the human gastrin gene were introduced into a plasmid vector downstream from the chloramphenicol acetyl transferase (CAT) gene and upstream from the SV40 early poly(A) site, specific sequences could be defined by deletion analysis as being capable of preventing the processing of pre-RNAs at the SV40 site (Sato *et al.*, 1986). These sequences included a stretch of 32 nts including 27 T and 5 A residues, beginning 190 nts downstream from the cleavage/polyadenylation site. Since pre-RNA 3' termini could be mapped to this region using S1 nuclease analysis, the (A+T)- rich sequence was designated a POLII terminator, though the possibility that it signals a processing event has not yet been excluded. Similar results have been obtained using a 1008 nt. fragment from the 3' end of the Ad2 major late transcription unit (Dressler and Fraser, 1987). In this case *in vitro* nuclear run-on was used to show that transcriptional termination, rather than transcript cleavage, was responsible for the absence of transcripts extending as far as the downstream SV40 poly(A) site.

Pulse-labelling studies have shown that transcription of the mouse β_{major}-globin gene terminates heterogeneously within the region between 700 and 2000 nts downstream from the poly(A) site (Citron *et al.*, 1984). In order to generate large amounts of nuclear RNA with which to assay the termination effect, recombinant adenoviruses were constructed which contained 3' portions of the β_{major}-globin gene within the Ad5 E1a transcription unit (Falck-Pedersen *et al.*, 1985). Pulse-labelling experiments performed both on whole cells and on isolated nuclei after infection with the reconstructed viruses indicated that E1a

transcription through to downstream sequences was prevented by inclusion of the β_{major}-globin poly(A) site and an additional 1395 nts downstream. Again, the putative terminator is rather (A+T)-rich, containing in addition three copies of the AATAAA hexanucleotide, as well as a sequence resembling a prokaryotic rho-independent termination site. The fact that there is no extensive sequence homology, other than a general (A+T)-richness, between the sea urchin H2A, human gastrin and mouse β_{major}-globin terminators strengthens the view that the phenomena being observed are neither straightforward nor universal.

While it is reasonable to suggest that termination of transcription is an important means through which expression of a given gene is prevented from interfering with the expression of adjacent, downstream genes, experimental results bearing on this point appear to be contradictory. In the case of the avian leukosis virus, insertional activation by the 3' long terminal repeat (LTR) appears to depend upon inactivity of the 5' LTR. This clearly suggests that transcriptional readthrough from the upstream element impairs the function of the 3' promoter (Varmus and Swanstrom, 1982). Promoter exclusion of this sort has been demonstrated in both prokaryotic and eukaryotic model systems. In the first, transcription initiated at an upstream P_L promoter continued through a downstream *gal* promoter, whose activity was found to be severely reduced (Adhya and Gottesman, 1982). Similarly, transcriptional readthrough from an upstream gene has been shown to be highly detrimental to the expression of an α-globin gene (Proudfoot, 1986). This negative effect could be relieved by the insertion between the two tandemly arranged genes of a terminator from either the sea urchin H2A or the mouse β_{major}-globin gene. In contrast, insertion of the mouse β_{major}-globin terminator into the Ad5 E1a gene had a negative effect on expression of the downstream E1b gene (Falck-Pedersen *et al.*, 1985). A possible explanation of this seemingly anomalous result is that efficient initiation of E1b transcription might depend upon prior termination at the 3' end of the E1a transcription unit; termination at this site would be prevented by insertion of a functional terminator

upstream. Such a close association between terminator and promoter elements would not be without precedent: the POLI termination region in the ribosomal RNA (rRNA) gene spacer constitutes part of the promoter for the next downstream rRNA gene in both mouse (Henderson and Sollner-Webb, 1986; Grummt *et al.*, 1986) and Xenopus (McStay and Reeder, 1986). This arrangement might serve to provide a high local polymerase concentration for the downstream promoter of such tandemly arranged genes, without recourse to truly polycistronic transcription, which does not seem to be a strategy favoured by the eukaryotic cell. Thus, although unimpeded readthrough transcription from upstream can impair promoter function, the latter might be potentiated by the provision of a closely linked terminator upstream.

Transcription termination and 3' processing.

The process of mRNA 3' end formation in yeast appears to lie, in evolutionary terms, midway between the archetypal prokaryotic and eukaryotic mechanisms, for although yeast mRNAs are polyadenylated, their 3' ends appear to be defined by transcriptional termination, rather than endonucleolytic cleavage (Zaret and Sherman, 1982, 1984). A less intimate association of polyadenylation with transcriptional termination may have been conserved in certain higher eukaryotic systems. For example, the mouse β_{major}-globin terminator was found to be functional only when linked to its poly(A) site, in the normal transcriptional orientation (Falck-Pedersen *et al.*, 1985). Corroborative evidence has come from study of a human thalassaemic α-2 globin gene which has a non-functional poly(A) site (Whitelaw and Proudfoot, 1986). Transcripts derived from this gene fail to terminate within the normal region, some 600 nts beyond the poly(A) signal, again suggesting an involvement of a functional processing site in correct termination. For Py, termination of early region transcripts, which use an efficient poly(A) site, appears to occur shortly after traversal of this site at early times in infection; late region transcripts, whose poly(A) site is less efficiently recognised, extend many times around the genome,

being only inefficiently terminated (Acheson and Mieville, 1978; Acheson, 1984). One possible mechanism, taking into account the apparent dependence of termination efficiency upon processing efficiency, would be that after cleavage, the downstream segment of the pre-RNA is rapidly degraded in a 5' to 3' direction until the transcribing POLII complex is reached, resulting in a termination event not necessarily specified by the primary sequence of the transcript. Such a mechanism does not, however, account for termination at the 3' end of the sea urchin H2A gene, which is independent of RNA processing (Birchmeier *et al.*, 1984), or for the apparent termination within both the H2A terminator and the human gastrin element when positioned 5' to a poly(A) site (Johnson *et al.*, 1986; Sato *et al.*, 1985). It seems, then, that a number of alternative mechanisms for transcription termination may be employed by the higher eukaryotic cell, and that only in some instances is there a direct involvement of polyadenylation.

Differential Polyadenylation and Termination of Cellular Transcripts.

In cases where a cellular gene has more than one functional poly(A) site, differential use of these sites can be used to generate mRNAs with different 3' sequences, and in some cases this results in production of different protein products. A simple example is that of the mouse α-amylase gene, whose primary transcripts use either one of two poly(A) sites positioned 237 nts apart in the 3' untranslated region of the gene (Tosi *et al.*, 1981). Similarly, the mouse dihyrdofolate reductase gene has four poly(A) sites within an 850 nt. region, again downstream from the translational termination codon (Setzer *et al.*, 1980). In the latter case, utilisation of the poly(A) sites has been shown to be specifically stimulated in actively growing cells (Kaufman and Sharp, 1983), but this stimulation appears to be equal for all four sites, and so the presence of more than one site is of no known biological consequence.

In contrast, alternative pathways of 3' processing have been shown to alter the protein products of the mouse IgM and rat calcitonin genes, as described in an earlier section. It

appears that the splicing pathway may be determined by the choice of polyadenylation site; deletions which remove the poly(A) site used by secreted μmRNA restore the production of mRNAs encoding membrane-bound IgM (Danner and Leder, 1985). The selection of calcitonin versus CGRP and secreted versus membrane-bound IgM poly(A) sites does not seem to be based upon differential transcriptional termination. Nuclear transcription was shown to extend beyond the sixth exon of the rat calcitonin gene in both thyroid C and brain cells (Amara *et al.*, 1984). Similarly, Yuan and Tucker (1984) reported that no decrease in RNA synthesis was observed between the first and second poly(A) sites in the μ gene in B cells, regardless of their developmental stage. Transcriptional termination may, however, be involved in a further aspect of immunoglobulin gene regulation. Early B cells may express IgD as well as IgM at their cell surface, and the δmRNA is derived by the use of downstream exons on the same primary transcript which gives rise to both forms of μmRNA. Such cells exhibit a constant, relatively low level of RNA synthesis across the heavy chain region encompassing all of the μ and δ exons (Mather *et al.*, 1984; Yuan and Tucker, 1984). Upon antigenic or mitogenic stimulation of such cells, there is a large increase in the rate of RNA synthesis within the region of the upstream μ exons, but the transcription of the downstream, δ exons remains at the initial, low level. This increase in transcription is accompanied by preferential use of the most promoter-proximal poly(A) site, and the generation of μmRNA, as described earlier. Thus it would appear that the synthesis of δmRNA is disfavoured in mature B cells not only by selection of the μ poly(A) site, but also by termination of the majority of nuclear transcripts in the region between the μ and δ exons. It is conceivable, in the light of evidence for the involvement of prior polyadenylation in transcription termination discussed earlier, that the termination seen in the heavy chain gene is simply the result of selection of the upstream poly(A) site.

Results obtained using the c-*myc* gene, in contrast, appear to implicate a termination mechanism unrelated to polyadenylation in the regulation of gene expression (Bentley and Groudine, 1986). During *in vitro* differentiation of human promyelocytic leukaemia cells

into granulocytes, there appears to be no change in the rate of transcriptional initiation at the c-*myc* promoters. In the undifferentiated cells there was an approximately 3-fold molar excess of exon 1 transcription over exon 2, and this excess increased to about 15-fold after differentiation. This correlated with a 10-fold reduction in the steady-state level of c-*myc* mRNA. In this case, then, a transcriptional block appears to be largely responsible for the observed decrease in absolute mRNA level. This is in contrast to the IgM/IgD control, where a large increase in the rate of transcription initiation is necessary to achieve the shift to favour μmRNA production.

T-AG FUNCTIONS IN THE PY AND SV40 LYTIC CYCLE

The existence of two SV40 and three Py predominant early proteins (Fig. 1.1) has been explained in terms of the DNA sequences of the viral genomes and cDNAs (Tooze, 1981; Treisman *et al.*, 1981a). These predictions correlated well with the results of *in vitro* translation of viral mRNA and immunoprecipitation using sera raised against virally-induced tumours. The large T-Ags, whose size has been estimated at between 81 and 100 kDa, reside mainly in the nucleus, although about 5% of SV40 large T-Ag may be associated with the plasma membrane lamina, possibly as the result of acylation (Schaffhausen, 1983). Both SV40 and Py large T-Ags are phosphorylated, in the case of the SV40 protein at two clusters of serine and threonine residues, one near each terminus of the polypeptide (reviewed by Salzman, 1986). As well as acylation and phosphorylation, other post-translational modifications of large T-Ag, including ADP-ribosylation, adenylation, methylamination, glycosylation and acetylation of the N-terminal methionine residue, may be involved in alteration of the proteins' cellular location or diverse activities. The capacity of the proteins to oligomerise (Bradley *et al*, 1982) may also relate to these activities. An N-terminal sequence of the SV40 large T-Ag is required for localisation of the protein in the nucleus (Lanford and Butel, 1980a, b), and fusion of amino acids 126-132 to a normally cytoplasmic protein resulted in the majority of the fused peptide being transported to the nucleus (Kalderon *et al.*, 1984a, b). Similarly, a

Py mutant large T-Ag missing amino acids 191-209 was defective in its ability to enter the nucleus (Nilsson and Magnusson, 1984). Relative to the large T-Ags, the small and middle T-Ags are only relatively poorly characterised. The 17-22 kDa small T-Ags are found both in the nuclear and in the cytoplasmic cellular compartments. They do not appear to be phosphorylated, but, like the large T-Ags, their N-terminal methionine residues may be methylated (Paucha *et al.*, 1978). The 55 kDa middle T-Ag, unique to Py, was originally found to be located in the plasma membrane fraction (Ito *et al.*, 1977). Subsequent investigation has suggested that the protein is also found in the perinuclear region and the cytoplasm (Zhu *et al.*, 1984). Middle T-Ag is partly phosphorylated, and while the protein is distributed approximately equally between the plasma membrane and the rest of the cell, 80-90% of the *in vivo* phosphate labelled middle T-Ag is in the membrane fraction (Segawa and Ito, 1982). As with large T-Ag, the majority of the phosphorylation appears to be on serine and threonine residues.

Aside from the function of large T-Ag in repressing transcription from SV40 and Py early promoters, a number of other T-Ag functions have been described which go some way towards an explanation of the ways in which these proteins control the progression of productive viral growth. Large T-Ag induces the transcription of a number of cellular genes, including the rRNA (Learned *et al.*, 1983), heat shock (Khandjian and Turler, 1983) and thymidine kinase genes (Postel and Levine, 1976; Schutzbank *et al.*, 1982). The latter may have an important secondary effect on the stimulation of viral and cellular DNA replication. Solution hybridisation data obtained using cytoplasmic RNAs from a human fibroblast line and an SV40-transformed derivative showed that the majority of mRNA sequences were common to both cell types (Williams *et al.*, 1977). However, about 3% of the transformed cell mRNAs were not present in the normal cell, and about 5% of normal cell mRNAs were repressed in the transformed cell. The identity of genes specifically activated by the presence of SV40 large T-Ag has been investigated by differential screening of cDNA libraries and molecular cloning (Schutzbank *et al.*, 1982;

Scott *et al.*, 1983). Some such genes, for example one which encodes a Qa/T1a class I major histocompatibility complex antigen, are also activated in mouse cells transformed by other agents. There were some clones whose activation, in contrast, appeared to be more characteristic of SV40 transformation, including an mRNA encoding a 58 kDa mitochondrial protein (Zuckerman *et al.*, 1984). Stimulation of cellular gene expression may be a consequence of large T-Ag binding to cellular DNA (Oren *et al.*, 1980). A fraction of SV40 large T-Ag forms a complex with the cellular protein p53 (Lane and Crawford, 1979). The proteins are present in equimolar amounts in the complex (Freed *et al.*, 1983), which recruits all of the available p53 (Greenspan and Carroll, 1981). Other properties of the large T-Ag polypeptides which may relate to their effects *in vivo* include ATPase (Clark *et al.*, 1984), protein kinase (Tjian and Robbins, 1979) and nucleotidyl transferase (Bradley *et al.*, 1984) activities, though these are, as yet, only poorly characterised. The direct role of large T-Ag proteins in viral DNA replication and the stimulation of late transcription will be discussed in later sections.

The requirement for small and/or middle T-Ags in the Py lytic cycle in animals has been investigated by McCance (1981). Host range mutants, bearing mutations in the sequences spliced out of the mRNA for large T-Ag but retained in the mRNAs for small and middle T-Ags, were found to grow very poorly in mice, with maximum titres 4 to 5 logs below those for wild type virus. Host range mutants of this type were found to be defective in VP1 phosphorylation and virus assembly (Garcea *et al.*, 1985). A further study implicated small, rather than middle, T-Ag in the cooperation with large T-Ag required for viral growth in mouse 3T6 cells (Templeton *et al.*, 1986), and Py small T-Ag has also been found to enhance viral DNA replication in these cells (Berger and Wintersberger, 1986). These observations aside, the roles of small and middle T-Ags in the productive cycle remain rather nebulous. The SV40 small T-Ag induces the dissolution of actin cables (Bikel *et al.*, 1986), affects other cellular structures, including centrioles (Shyamala *et al.*, 1982), and binds to host proteins (Rundell, 1982). Much of what is

known about Py middle T-Ag function has come from studies of transformed cells, but one observation which may be relavent to the productive cycle is that middle T-Ag alone is able to stimulate transcription from a mouse H-2 class I promoter (Majello *et al.*, 1985). It is not known if this and/or other genes are similarly affected by middle T-Ag during infection. Middle T-Ag associates with the product of the c-*src* proto-oncogene (Courtneidge and Smith, 1983). A possible role for this complex in cell transformation is discussed in a later section, but as with the complex between SV40 large T-Ag and p53, there is no proven involvement of this complex in the viral lytic cycle. The most attractive explanation is that such complexes represent components of the pathways through which these viruses are able to stimulate entry of their host cells into the mitotic cycle.

VIRAL DNA REPLICATION

Py and SV40 viral DNA synthesis is first detectable 12-15 hours after infection of permissive cells, at which stage cellular DNA synthesis is also induced. Initiation of SV40 or Py DNA replication involves at least three basic components: host replication factors, including those which distinguish permissive from non-permissive cells; the viral origin of replication (*ori*) and the homologous large T-Ag, which binds to specific DNA sequences in and around *ori*. The minimum size of the SV40 *ori*, 64 bp between nts 5208 and 30, was defined by mutations on either side which still allowed DNA replication, albeit with reduced efficiency (reviewed in Salzman, 1986). This region contains a perfect 27 bp dyad symmetry adjacent to a 17 bp sequence of (A+T), and includes large T-Ag binding site II. This apparent similarity to prokaryotic rho-independent transcription termination sites has been related to the capacity of the SV40 *ori* to terminate POLII transcription *in vitro* (Grass *et al.*, 1987), but the significance of this finding is not yet clear. Similarly the Py *ori* core, located between nts 5265 and 65, contains a dyad symmetry element adjacent to a 14 bp (A+T)-rich sequence, but in this case the *ori* contains none of the strong large T-Ag binding sites (Fig. 1.2). The Py *ori* requires, in addition to the core

sequence, sequences contained within the viral enhancer in order to function, and a similar link between replication and transcription functions may exist for SV40. The minimal sequences for Py *ori* activity consist of the core sequence plus either of two regions defined by deletion analysis, which closely correspond to the enhancer domains A and B described earlier (Luthman *et al.*, 1982; Katinka and Yaniv, 1983; Muller *et al.*, 1983). The finding that other enhancers can substitute for the Py function in replication has raised the possibility that Py DNA replication may be transcriptionally activated (de Villiers *et al.*, 1984).

Permissive cells appear to provide factors required for replication (rather than non-permissive cells providing factors which repress replication), as judged by cell fusion experiments (Botchan *et al.*, 1980, and references therein). The permissive factors could, in theory, control viral DNA replication at a number of levels: they could regulate the level of large T-Ag at the level of its transcription or translation; they could be responsible for post-translational modification of large T-Ag to give an active form, or they could participate directly in the initiation of DNA synthesis. A simple explanation of host cell control over viral DNA replication through regulation of large T-Ag synthesis does not account for all of the pertinent experimental results. For example, levels of large T-Ag undetectable by immunofluorescence were sufficient for SV40 DNA synthesis in permissive cells (Lanford and Butel, 1980). In addition, certain Py enhancer mutants selected for growth on embryonal carcinoma (EC) cells did not replicate in other, independently isolated EC cell lines, despite the fact that the mutant enhancer was active in both cell types (Melin *et al.*, 1985a, b, and references therein). Furthermore, substitution of the Py enhancer for that of SV40 allowed efficient large T-Ag production but not SV40 DNA replication in mouse cells (Weber *et al.*, 1984). Apparently, then, viral DNA synthesis is initiated in permissive cells because these cells contain factors which cause an active replication complex to form between the large T-Ag and its homologous *ori* core region. In the case of Py, the nature of the enhancer element required for *ori* function can

also impart cell-type specificity, at least within the mouse. Py small T-Ag stimulates the replication of viral genomes in mouse 3T6 cells, and the extent of this stimulation varies in different cell lines (Berger and Wintersberger, 1986). Py and SV40 mutants defective in small T-Ag expression had previously been shown to grow in certain, though not in most, cell lines (Shenk *et al.*, 1986; Silver *et al.*, 1978). These results suggest that a cellular factor required for replication may be limiting in some cells in the absence of small T-Ag. The first identification of cellular factors which contribute to the definition of viral host range has recently been facilitated by the development of *in vitro* replication systems for SV40 (Li and Kelley, 1984) and Py (Murakami *et al.*, 1986a). SV40 large T-Ag-mediated DNA replication occurs in mouse cell extracts only when they are supplemented with purified human or monkey polymerase-α primase (Murakami *et al.*, 1986b). Conversely, Py DNA will only replicate in human cell extracts in the presence of purified mouse polymerase-α primase complex.

The involvement of large T-Ag was first suggested by the results of temperature shift experiments using ts-A mutant viruses (Tegtmeyer, 1972; Francke and Eckhart, 1973). These studies showed that large T-Ag was required for initiation, rather than elongation or termination of each round of DNA synthesis. SV40 tsA mutants defective in viral replication had large T-Ags with reduced ability to bind viral DNA in a specific manner, and the capacity of the protein to oligomerise also appeared to be impaired. In many cases the replication defect was also seen as a transformation defect, although induction of cellular DNA synthesis was not necessarily affected (Setlow *et al.*, 1980). A specific role for SV40 large T-Ag in the initiation of viral DNA synthesis has been proposed by Wold *et al.* (1987), whose studies suggested that the protein binds to the viral *ori* and locally unwinds the DNA. This *ori*-dependent unwinding reaction is presumably a prerequisite for subsequent priming and elongation steps. Other viral gene products may be involved in this mitogenic function, as Py host range mutants defective in small and middle T-Ag production cause a greatly reduced induction of cellular DNA synthesis (Turler and

Salomon, 1985). Products of the SV40 early region stimulated cellular DNA synthesis in hamster cells temperature sensitive for DNA replication at both permissive and non-permissive temperatures (Floros *et al.*, 1981). This result suggested that SV40 T-Ags bypass a requirement for factors involved in the progression from G1 to S phase.

A number of lines of evidence suggest that replication proceeds bidirectionally from the unique *ori*. Cleavage of replication intermediates with restriction enzymes, followed by electron microscopy, and measuring the presence of radioactivity in restriction fragments after pulse-labelling have both shown that the two replication forks progress at equal rates, and converge on a point diametrically opposite the *ori* (reviewed in Tooze, 1981). Rolling circle intermediates have also been detected (Bjursell, 1978) but appear to be prevalent only at late times in infection of permissive cells, or during inefficient replication in non-permissive cells, suggesting that these forms are probably aberrant. For the bulk of replicating molecules, the parental strands appear to remain covalently closed during replication, and a topoisomerase activity is thought to release the torsional strain that accumulates as the DNA helix is unwound (Champoux and Been, 1980). Chain elongation is probably continuous on the "forward" arm of each replication fork, and discontinuous on the "retrograde" arm, involving the repeated initiation of Okazaki fragments (reviewed in Salzman, 1986). The position of termination appears to be specified by the meeting point of the two replication forks, rather than by a particular DNA sequence. Thus defective viral molecules are replicated efficiently, even though they may lack the region where termination normally occurs (Martin *et al.*, 1975; Griffin and Fried, 1975).

LATE TRANSCRIPTION

At late times in infection transcription of the late (L) DNA strand of Py or SV40 is at least ten times as efficient as transcription of the early (E) DNA strand, as measured by analysis of pulse-labelled RNA synthesised *in vivo* (Laub and Aloni, 1975; Flavell and Kamen, 1977) or by analysis of unlabelled RNA synthesised *in vitro* by viral

transcription complexes (Ferdinand *et al.*, 1977; Condit *et al.*, 1977). The bulk of late RNA is probably synthesised on newly replicated viral templates. The high levels of late transcripts reflect the requirement late in infection for large amounts of viral capsid proteins; of the order of 10^8 VP1 molecules would be required to encapsidate the 2×10^5 viral DNA molecules produced in each infected cell (Tooze, 1981).

Differential splicing is used to produce the three Py and two SV40 late mRNA size classes, which encode the three viral capsid proteins. In each case the minor proteins, VP2 and VP3, are translated in the same reading frame, using alternative initiation codons, with VP1 being translated in a different phase from a partially overlapping reading frame. The different splices employed in the late region of each virus are represented diagrammatically in figure 1.1. The late RNAs have unusually heterogeneous 5' ends. For SV40, when the 5' ends of nascent late RNAs were determined by S1 nuclease analysis, three species, initiated at nts 325, 260 and 195 were detected (Lycan and Danna, 1983), with transcripts beginning at nt. 325 accounting for the majority of the signal. These results confirmed the earlier findings of Canaani *et al.* (1979) and Haegeman and Fiers (1978), who analysed SV40 late RNA cap structures and found the same major start site. The late transcripts appeared to be initiated almost exclusively at adenine residues. For Py, 90% of the late transcription start sites are at one of the purine residues in a 25 bp region between nts 5077 and 5101 (Cowie *et al.*, 1981), although late transcription *in vitro* used only the site at nt. 5129, which is located 32 bp downstream from a TATA box consensus (Jat *et al.*, 1982). Py late nuclear RNA contains tandemly repeated transcripts of the entire L DNA strand (Birg *et al.*, 1977; Acheson, 1978), only a minority of which is polyadenylated (Lev *et al.*, 1979; Acheson, 1984). The apparent inefficiency of Py late polyadenylation has been proposed as the reason for this inefficiency of transcriptional termination by Lanoix *et al.* (1986). Their results suggested that inefficient polyadenylation and termination result from the limited availability late in infection of cellular factor(s) required for these processes, rather than inherent inefficiency of the late

poly(A) site. When RNAse T1 digests of late Py mRNAs were examined, the region between 65 and 18 nts upstream from the VP2 initiation codon was found to be present in three- to five-fold molar excess. This "leader" sequence is reiterated at the 5' ends of the late mRNAs, an arrangement presumed to arise through post-transcriptional processing of the long primary transcripts (Legon *et al.*, 1979; Treisman, 1980). For SV40, transcription continues past the late poly(A) site before terminating, apparently at random (Ford and Hsu, 1978). That SV40 late transcription termination appears to be more efficient than that of Py may reflect a greater abundance of factor(s) necessary for polyadenylation and/or termination in the monkey cells which support SV40 growth. In the majority of cases, the SV40 late poly(A) site functions after the first transit of the polymerase, but a few transcription complexes probably complete more than one circuit of the entire L DNA strand, producing a tandemly repeated RNA which is then processed in the manner of the Py giant transcripts to yield variant RNAs with 5' leader sequences containing a tandem repeat (Reddy *et al.*, 1978).

The opposed arrangement of the early and late transcription units in Py and SV40 leads one to suggest that transcriptional, and indeed post-transcriptional, interference between the two units might complicate viral gene expression. For example, the traversal of the Py early promoter region by L DNA strand transcription complexes engaged in synthesis of the giant nuclear RNAs might be expected to interfere with the assembly of transcription complexes on the E DNA strand. In addition, the possible steric hindrance experienced by POLII complexes attempting to travel in opposite directions along the viral template might cause premature termination of transcription of the E DNA strand at late times in infection. The problem of simultaneous transcription of both template DNA strands has been investigated in prokaryotic systems, using different pairs of opposed promoters. Overlapping transcription was found to reduce the efficiency of RNA production, especially when the promoter had an intrinsically low activity (Ward and Murray, 1979; Horowitz and Platt, 1982; Lloubes *et al.*, 1986). This general conclusion

has been confirmed in one eukaryotic case, that of the gene CYC1 which encodes iso-1-cytochrome C in yeast. The cyc1-512 mutation is a 38 bp deletion which appears to prevent normal transcription termination in this gene. This results in the overlapping, convergent transcription of CYC1 and an adjacent gene, which in turn causes a marked reduction in the mRNA levels of both genes (Zaret and Sherman, 1982). It has not yet been possible to establish whether or not converging transcription on both DNA strands occurs simultaneously for a given Py or SV40 template molecule, due to the inavailability of experimental tools with which to examine the phenomenon. While the amount of early RNA synthesised late in infection represents only a small proportion of the total viral transcription, one could conceive that, with the majority of templates exclusively directing L DNA strand transcription, a small proportion of the templates retain the early-type promoter-transcription factor complex. This model would not, then, involve direct transcriptional interference between the two viral genes.

A number of lines of evidence suggest that the efficiency of nuclear-cytoplasmic transport and/or translation of eukaryotic transcripts present in double-stranded RNA complexes is markedly reduced (Harland and Weintraub, 1985; Kim and Wold, 1985; Melton, 1985). There are two points at which such effects could come into play in the late phase of Py and SV40 gene expression. Firstly, the abundant giant Py L strand nuclear transcripts, being in considerable excess over nuclear E strand pre-RNAs, might be capable of driving inter-molecular hybridisation, with consequent reduction of E strand expression. Secondly, the mature, predominant early and late mRNAs of both viruses overlap at their 3' ends, by 88nt for SV40 and by approximately 41 nts for Py. Thus it is possible that the excess of L strand mRNAs could interfere with residual early gene expression late in infection, through the hybridisation of the complementary 3' mRNA sequences. Py early region transcripts polyadenylated at the alternative site at the 3' end of the middle T-Ag coding sequences would be susceptible to "antisense" inhibition by the first, but not by the second mechanism, since there is no 3' overlap between these

transcripts and mature late region mRNAs.

The sequence elements which define the late promoters have not yet been thoroughly characterised. The significance of the TATA box sequence TAATTAAAAG in the late region of the Py (A2 strain), just upstream from the predominant transcriptional start sites, is uncertain; mutation of the sequence to TAACTAAAAG in Py strain WSWT does not seem to have had any effect on virus viability (Deininger *et al.*, 1979). In the case of SV40 late transcription there is no TATA box-like sequence at an appropriate distance from the mRNA 5' ends. Whereas the 21 bp repeats are essential for both early and late transcription *in vitro* (Rio and Tjian, 1984; Vigneron *et al.*, 1984), *in vivo* transcription of the SV40 late genes was only partially suppressed by deletion of the repeats (Fromm and Berg, 1982). Furthermore, after transfection of human HeLa cells with plasmids containing the SV40 enhancer and late promoter sequences, late transcription occurs efficiently whether or not the 21 bp repeats are present (Omilli *et al.*, 1986).

Between the early and late phases of infection, the ratio of L:E DNA strand transcription changes from about 1:5 to about 20:1. In this time interval, the overall rate of viral RNA synthesis increases dramatically, presumably largely as the result of the increased numbers of template molecules. For E strand transcription, this increase is about 50-fold, whereas for L strand transcription it is about 5,000-fold (Tooze, 1981). The observation that late region transcription is predominant only after viral DNA replication has led to a number of investigations of a possible cause-effect relationship between the two functions. However, the results of these studies clearly show that late transcription can proceed in the absence of DNA replication. For example, when inhibitors of DNA synthesis were used to block replication late in viral infection, the rate of L strand transcription remained unchanged (Manteuil and Girard, 1974). In recombinants which lack the SV40 *ori*, the late promoter can still function to initiate transcription from the authentic late start sites (Ernoult-Lange *et al.*, 1984).

Since large T-Ag is required for viral DNA replication, and hence for the increase in

template copy number at late times in infection, there is an indirect link between large T-Ag synthesis and the induction of L strand transcription. In addition, several studies have shown that late transcription is directly stimulated by this protein. During the early phase, before sufficient amounts of large T-Ag are made, L strand sequences are transcribed at a low rate, with a constitutively higher level of E strand transcription accounting for the 1:5 ratio of L:E strand RNAs. Temperature shift of cells infected with SV40 tsA mutants has been used to inactivate large T-Ag after the onset of viral DNA replication. Under these conditions, L strand RNAs continued to be synthesised at high rates (Cowan *et al.*, 1973; Khoury and May, 1977). This is despite the fact that large T-Ag is required for the initial increase in L strand transcription (Brady *et al.*, 1984; Keller and Alwine, 1984). Thus the rate of L strand transcription increases dramatically when a threshold level of large T-Ag is available, and this stimulation cannot be reversed by using thermolabile large T-Ags. Brady *et al.* (1985) have shown that large T-Ag-dependent stimulation of late transcription requires viral DNA sequences at T-Ag binding site II and within the enhancer. Further evidence for an involvement of large T-Ag binding to the viral template in the activation of L strand transcription has come from studies using transformed cells (Keller and Alwine, 1984). Cells producing altered large T-Ags with decreased DNA-binding capacity also have decreased late promoter activity. The possibility remains that large T-Ag is additionally involved in indirect stimulatory mechanisms, for example the modification or modulation of POLII or an as yet unidentified transcription factor.

VIRUS ASSEMBLY

A large proportion of the Py and SV40 DNA synthesised, perhaps 90%, is encapsidated (Garber *et al.*, 1978). Studies on assembly intermediates and their structural protein composition have suggested that the mechanism of assembly involves the gradual addition of viral structural proteins to viral chromatin complexes. The results of *in vitro* reassembly of Py capsomeres and chromatin into infectious virions support this general

view of assembly (Yuen and Consigli, 1982, and references therein).

The morphology of Py and SV40 viral particles deduced from electron micrographs suggested that the capsid was constructed from 12 pentameric and 60 hexameric capsomeres on an icosahedral surface lattice (Klug, 1965). However, the 60 hexavalent capsomeres are, in fact, pentamers that are structurally indistinguishable from the 12 pentavalent units (Rayment *et al.*, 1982). This finding suggested that nonequivalent bonding must occur between the molecules of VP1, and in turn that different forms of VP1 might have distinct roles in capsid formation. Underphosphorylation of the two most acidic VP1 species is accompanied by defective virion assembly in Py early region *hr-t* mutants (Garcea *et al.*, 1985). This result implies firstly that specific phosphorylation of VP1 sub-species is required for authentic virion assembly, and secondly that small and/or middle T-Ags are involved in this phosphorylation step. However, purified VP1 prepared by expression in *E. coli* was seen spontaneously to form first pentameric capsomeres and then entire capsid-like assemblies. This self-assembly of the unmodified VP1 implied that neither the post-translational modifications of VP1 nor the presence of VP2 and VP3 were essential for capsid formation (Salunke *et al.*, 1986).

The productive viral cycle is completed as cellular integrity is lost, and of the order of 10^5 progeny viruses are released. Mutants deficient in viral DNA synthesis and encapsidation also display a reduced cytopathic effect (Norkin, 1977), but the means by which this effect are elicited have not been extensively studied.

CELL TRANSFORMATION BY PY AND SV40

An earlier section summarised results which identified the tumorigenic potential of Py and SV40 injected into appropriate hosts. The development of cell culture systems in which the viruses induce a transformed phenotype of altered growth properties and morphology has been of fundamental importance in the approach to an understanding of this oncogenicity. Py can transform rodent cells *in vitro* to a phenotype indistinguishable from that of a Py-induced tumour cell derived *in vivo* (Vogt and Dulbecco, 1960). The

transformed cells gave rise to tumours when injected into susceptible animals, demonstrating the validity of the assay. Similar results were obtained using SV40 (Todaro and Green, 1964). Whether transformed *in vivo* or *in vitro*, the cells were invariably found to contain an expressed viral early region integrated into the cellular chromosome (Sambrook *et al.*, 1968; Birg *et al.*, 1979). Further experiments using animal systems, temperature-sensitive mutants and DNA transfection of cultured cells have clearly shown that early viral functions are sufficient both for tumour induction *in vivo* and cell transformation *in vitro* (Dulbecco and Eckhart, 1970; Tegtmeyer, 1975; Van der Eb *et al.*, 1979; Moore *et al.*, 1980). The results of experiments using thermolabile large T-Ag mutants clearly identified large T-Ag as the transforming gene of SV40 (Brugge and Butel, 1975; Kimura and Itagaki, 1975; Martin and Chou, 1975; Osborn and Weber, 1975; Tegtmeyer, 1975). In contrast, the Py large T-Ag, whilst required for the initiation of the transformed state both *in vitro* and *in vivo* (Fried, 1965, DiMayorca *et al.*, 1969), was found to be neither intact nor functional in a number of transformed cell lines (Fried *et al.*, 1986 (see appendix 2) and references therein). Furthermore, rodent cells stably expressing Py large T-Ag as their only viral gene product were otherwise phenotypically normal (Lania *et al.*, 1980), and Py-transformed cells express both small and middle T-Ags (Ito and Spurr, 1979). Thus Py large T-Ag is required for the initiation but not the maintenance of transformation. The identity of the Py transforming gene was eventually confirmed by Treisman *et al.* (1981), who showed that a cDNA encoding only the middle T-Ag is sufficient to transform established cell lines.

During the lytic cycle, the viral functions implicated in the stimulation of host cell mitosis are, for SV40, large T-Ag (Graessman *et al.*, 1981) and, for Py, functions other than large T-Ag (Fried, 1970). There is thus a good correlation between viral functions required for the induction of cell DNA synthesis in productive growth and viral transforming functions. This has led to the suggestion that continued production of these early viral gene products in the absence of viral replication causes the continuous mitotic

stimulation which partially characterises the transformed phenotype. The necessary failure of viral replication could result either from, in the case of a heterologous host cell, the absence of compatible host functions, or from a defect in the viral components necessary for replication. It is not yet clear which of the properties of Py middle T-Ag are responsible for transformation. It is possible that middle T-Ag *trans*-activates cellular genes normally involved in the control of mitosis, since *trans*-activation of cellular genes by middle T-Ag has been demonstrated (Majello *et al.*, 1985). An alternative mechanism through which middle T-Ag might act was suggested by the finding that immunoprecipitates of the viral protein have an associated tyrosine kinase activity (Smith *et al.*, 1979; Eckhart *et al.*, 1979; Schaffhausen and Benjamin, 1979). There appeared to be a correlation between the kinase activity, measured *in vitro* in immunoprecipitates from cells infected with Py mutants, and the transforming ability of these mutants. The kinase activity did not seem to be a property of the purified middle T-Ag protein (Walter *et al.*, 1982), but rather was thought to derive from the small proportion of the cellular c-*src* gene product with which a small sub-population of the middle T-Ag protein appears to complex (Courtneige and Smith, 1983, 1984). This discovery led to a theory of Py transformation based on modulation of the $pp60^{c\text{-}src}$ protooncogene product by middle T-Ag. In the process of *in vitro* phosphorylation, middle T-Ag itself becomes phosphorylated, predominantly at the tyrosine residue at position 315. However, phosphorylation at this site is not observed *in vivo* (Schaffhausen and Benjamin, 1981), and indeed there is no increase in the total amount of phosphotyrosine residues in Py-transformed cells (Sefton *et al.*, 1980), in contrast to the increase, thought to be due to the action of the viral *src* gene product, observed in cells transformed by Rous sarcoma virus. In addition, cells containing a kinase-positive complex between a truncated middle T-Ag and $pp60^{c\text{-}src}$ were shown to have a normal, untransformed phenotype (Wilson *et al.*, 1986). The involvement, if any, of the middle T-Ag-$pp60^{c\text{-}src}$ complex in Py transformation therefore awaits further study.

The roles in cellular immortalisation and transformation which, for Py, are divided between the large and middle T-Ags, are combined in the large T-Ag of SV40. It could be proposed that the phenotype of SV40-transformed cells results from cooperation between the nuclear form of large T-Ag, which binds to DNA and could regulate the expression of cellular genes (Scott *et al.*, 1983), and a cytoplasmic or membrane-bound form of the same protein, which could play a role analogous to that of the Py middle T-Ag. In support of this model we have the results of Lanford and Butel (1984), who showed that a mutant defective in transport of large T-Ag to the nucleus was still able to transform established rodent cells, but was impaired in transformation of primary cells. However, studies of SV40 mutants have shown that the DNA-binding, Viral DNA replication and ATPase activities of large T-Ag are dispensable for transformation (Cole *et al.*, 1986, and references therein). Considerable efforts have been made to establish a role in transformation for the complex which forms between a proportion of SV40 large T-Ag and a cellular 53 kDa protein, p53 (Lane and Crawford, 1979), but at this stage results remain equivocal.

The potential involvement of Py small T-Ag in transformation has been investigated by using a cDNA encoding only small T-Ag (Zhu *et al.*, 1984). Cultured cells expressing large amounts of this protein became rounded and easily detached from plastic culture dishes, a phenotype probably related to the ability of the small T-Ag to disrupt actin cables (Bikel *et al.*, 1986). Studies with SV40 suggested that small T-Ag provides a "facilitator" function in transformation. This function can be provided by some actively growing cells (Seif and Martin, 1979; Sompayrac and Danna, 1983), and can be dispensed with when the level of expression of large T-Ag is very high (Kriegler *et al.*, 1984).

Viral Transcripts in Transformed Cells.

Cells transformed by Py or SV40 continuously express viral information derived from the early region(s) of the integrated viral sequences. In the case of SV40, the requirement

for an intact large T-Ag gene dictates that the early mRNAs generated are usually indistinguishable from those produced during lytic infection, although in some lines transcripts reading through into viral sequences from cellular promoters have been detected (Sambrook *et al.*, 1980). In general, such host-viral fusion transcripts would be expected to impair the SV40 early promoter function, and they may be selected against during transformation. For Py, transformed rat cells containing free viral genomes also contain early and late region mRNAs identical to those produced in lytic growth. The relative amount of late RNA was seen to decrease as freshly transformed rat cell lines were passaged in culture (Kamen *et al.*, 1979), correlating with the gradual disappearance of free Py genomes and presumably reflecting selection against cells with a high probability of induction. In contrast, Py-transformed mouse cells, being permissive for viral replication, are never found to contain free viral genomes. Although Py large T-Ag is required for the initiation of transformation, it appears to be selected against upon continued growth of transformed cells either *in vitro* (Basilico *et al.*, 1979a, b) or *in vivo* (Lania *et al.*, 1981). As a consequence, Py-transformed lines contain early sequences either deleted or otherwise mutated in the region unique to the coding frame for large T-Ag, while the selection for maintenance of an intact middle T-Ag gene is reflected in the retention of an unperturbed promoter-proximal half of the early region.

In lines where the large T-Ag sequences are disrupted by point mutation (Hayday *et al.*, 1983) or by deletion-insertion (Ruley and Fried, 1983) the nuclear viral transcripts are predominantly cleaved and polyadenylated at the major early poly(A) site. Other strategies for polyadenylation must be employed in the majority of Py-transformed lines, where disruption of the large T-Ag reading frame occurs through truncation of early region sequences, as a result of recombination with cellular DNA. In these lines the major early poly(A) site is invariably lost, and primary transcripts initiated at the viral early promoter extend into the downstream cellular sequences. When these pre-RNAs are cleaved and polyadenylated, the site of polyadenylation is frequently in the downstream

cellular sequences (Ruley *et al.*, 1982; Fenton and Basilico, 1981; B. Davies, personal communication). The frequency with which functional poly(A) signals will be found in the vicinity of random integration sites was estimated by Santangelo and Cole (1983). It was found that 56 of 67 randomly isolated fragments of African Green Monkey DNA, between 3.5 and 18 kb in length, were able to restore expression of an HSV thymidine kinase gene lacking its poly(A) site, when ligated downstream of the gene. It is perhaps not surprising, then, that in what is essentially an analogous *in vivo* assay, many random cellular sequences are able to restore efficient polyadenylation to the typically truncated Py early region following integration.

An alternative strategy for early mRNA polyadenylation is exhibited in the lines tsc3T3-2 and 53-Rat (Kamen *et al.*, 1979). In both lines the transcriptionally active early region lacks C-terminal large T-Ag sequences and their associated major poly(A) site. Just as in other transformed lines containing 3'-truncated early regions, hybrid nuclear transcripts initiated at the viral early cap sites extend into the cellular sequences downstream. Unlike other transformed cell pre-RNAs, the hybrid nuclear transcripts of tsc3T3-2 and 53-Rat are cleaved and polyadenylated almost exclusively at the alternative early poly(A) site (Kamen *et al.*, 1979). Furthermore, the amount of Py mRNA produced in these lines is of the same order as that found in lines wherein the Py major early or fortuitous cellular poly(A) sites are used (Kamen *et al.*, 1979 and this study). The 3'-truncated Py integrants present in various cell lines thus differ markedly in their 3' processing. Since the 5' expression sequences present in each line are identical, one must conclude that the differences in 3' processing reflect the different cellular locations into which each viral transcription unit has integrated.

In general, the pattern of splicing of the early region transcripts in Py-transformed cells has been reported to resemble that found in lytically infected cells (Kamen *et al.*, 1979; Fenton and Basilico, 1981). One notable exception was the finding of polyadenylated mRNAs with 3' ends of viral sequence in the region of nt. 1220 (Kamen

et al., 1979) in certain lines. These mRNAs were interpreted to be spliced derivatives of viral-host fusion pre-RNAs, the apparent 3' end indicating the point where viral sequences had been spliced to host sequences.

Viral Integration.

Permanent transformation of a cell after infection with SV40 or Py or after transfection with viral DNA is a rare event, though early after infection, a large proportion of non- or semi-permissive cells transiently acquire some of the properties associated with transformation (Stoker and Dulbecco, 1969). The limiting step in the transition from an abortively to a permanently transformed cell would seem to be the integration of viral sequences to permit continuous expression of the viral transforming functions.

For retroviruses, the integration of proviral sequences involves the generation of a direct repeat of a short cellular sequence on either side of an intact single copy of the viral DNA. No cellular sequences are lost as a consequence of retroviral integration, and in the cases of mouse mammary tumour virus and avian leukaemia virus the transformed phenotype is caused by integration within defined regions of the cellular genome and the subsequent activation of a cellular proto-oncogene (reviewed by Varmus, 1984). Integration of Py and SV40 displays none of these features. Although a number of transformed lines have been described which contain approximately one integrated viral genome equivalent per cell (Botchan *et al*, 1976; Fried *et al.*, 1986: see appendix 2), many contain head-to-tail tandem repeats of viral sequences, whose formation is apparently favoured by viral replication at the time of integration (Birg *et al.*, 1979; Campo *et al.*, 1978). However, the observation that such tandem repeats can be generated in fully non-permissive cells, or in cells infected with tsA mutants at the restrictive temperature, suggests that the repeats can be generated by homologous recombination between viral molecules, rather than by integration of rolling circle replicative intermediates. Although an active large T-Ag is not required for viral DNA

integration, the efficiency of transformation of rat cells by tsA mutant Py DNA was 20-fold higher at the permissive temperature than at the restrictive temperature (Della Valle *et al.*, 1981). Integrated viral DNA is frequently not colinear with the viral genome, being interrupted by direct repeats, deletions and deletion-insertions (Botchan *et al.*, 1980; Ruley and Fried, 1983). In some cases, short stretches of "filler" DNA, derived from neither viral nor directly adjacent host sequences, are found at viral-host joins (Hayday *et al.*, 1982; Stringer, 1982). In one such case, it was demonstrated that the filler DNA is an exact inverted duplication of a single copy cellular sequence found 650 bp from the viral-host join (Williams and Fried, 1986). After selection of transformed cell lines, the state of the integrated viral sequences has frequently been observed to change, usually from a complex to a simpler situation (Birg *et al.*, 1979; Hiscott *et al.*, 1981; Bender and Brockman, 1981).

Sequence determination at junction sites has not revealed extensive homologies between cellular and viral DNAs (Botchan *et al.*, 1980; Stringer, 1982; Fried *et al.*, 1986: see appendix 2), eliminating homologous recombination as a possible mechanism of integration. Despite this, short stretches of up to 5 bp of homology are frequently found at recombinant joins (Stringer, 1982; Ruley and Fried, 1983; Maruyama and Oda, 1984), and have been incorporated into a model for illegitimate recombination (Hasson *et al.*, 1984). In this model, endonucleolytic cleavage is followed by the pairing of free ends through the limited available nucleotide homology. Repair of the resulting heteroduplex would then generate the novel viral-cellular junction. In a further, exhaustive study by Roth and Wilson (1986), analysis of 199 independent junctional sequences was performed after introduction of a linearised plasmid with mismatched ends into monkey cells. Three mechanisms for end joining were defined: single strand, template-directed and postrepair ligations like those proposed by Hasson *et al.* (1984). The latter two mechanisms depend upon homologous pairing of up to 6 complementary nucleotides to position the junction, and all three mechanisms appeared to operate with similar efficiencies. While such mechanisms account for the structures of known viral-host

junctions, the means by which the DNAs are initially broken remain obscure. Bullock *et al.* (1985) found a correlation between junction breakpoints and topoisomerase I cleavage sites, but the sequences constituting such sites are only poorly, if at all, conserved.

One final common feature of Py and SV40 integration is that the cellular sequences on either side of the viral insert are not found adjacent to each other in untransformed cells (Hayday *et al.*, 1982; Stringer, 1982). This indicates that viral integration must have caused a deletion or translocation of the cellular DNA sequences at the integration site. This disruption of the cellular genetic material does not appear to be relavent to the phenomenon of transformation, however, since a haploid insertion of viral DNA is sufficient to elicit this phenomenon (Hayday *et al.*, 1982).

CHAPTER 2: THE POLYOMA VIRUS ALTERNATIVE EARLY POLYADENYLATION SITE DURING PRODUCTIVE INFECTION

The alternative Py early poly(A) site, located at the 3' end of the middle T-Ag coding sequences, could be important for productive viral infection *in vivo*. In this chapter I describe the investigation of the abundance, splicing pattern and detailed 3' structure of transcripts polyadenylated at the alternative site during productive Py infection of mouse 3T6 cells and in the transformed line 53-Rat. Northern blotting of mRNAs harvested at different times after infection gave a measure of the proportion of early Py mRNAs cleaved at the alternative site and also revealed their overall splicing pattern. The positions of cleavage/polyadenylation at the alternative site were investigated both by high-resolution 3' S1 nuclease mapping and by cDNA cloning. These results clarify the protein-coding potential of the truncated early mRNAs and show how the alternative poly(A) site could be used to favour the production of middle, and more especially small T-Ag at late times in infection, when these T-Ags may have specific roles.

OBJECTIVES

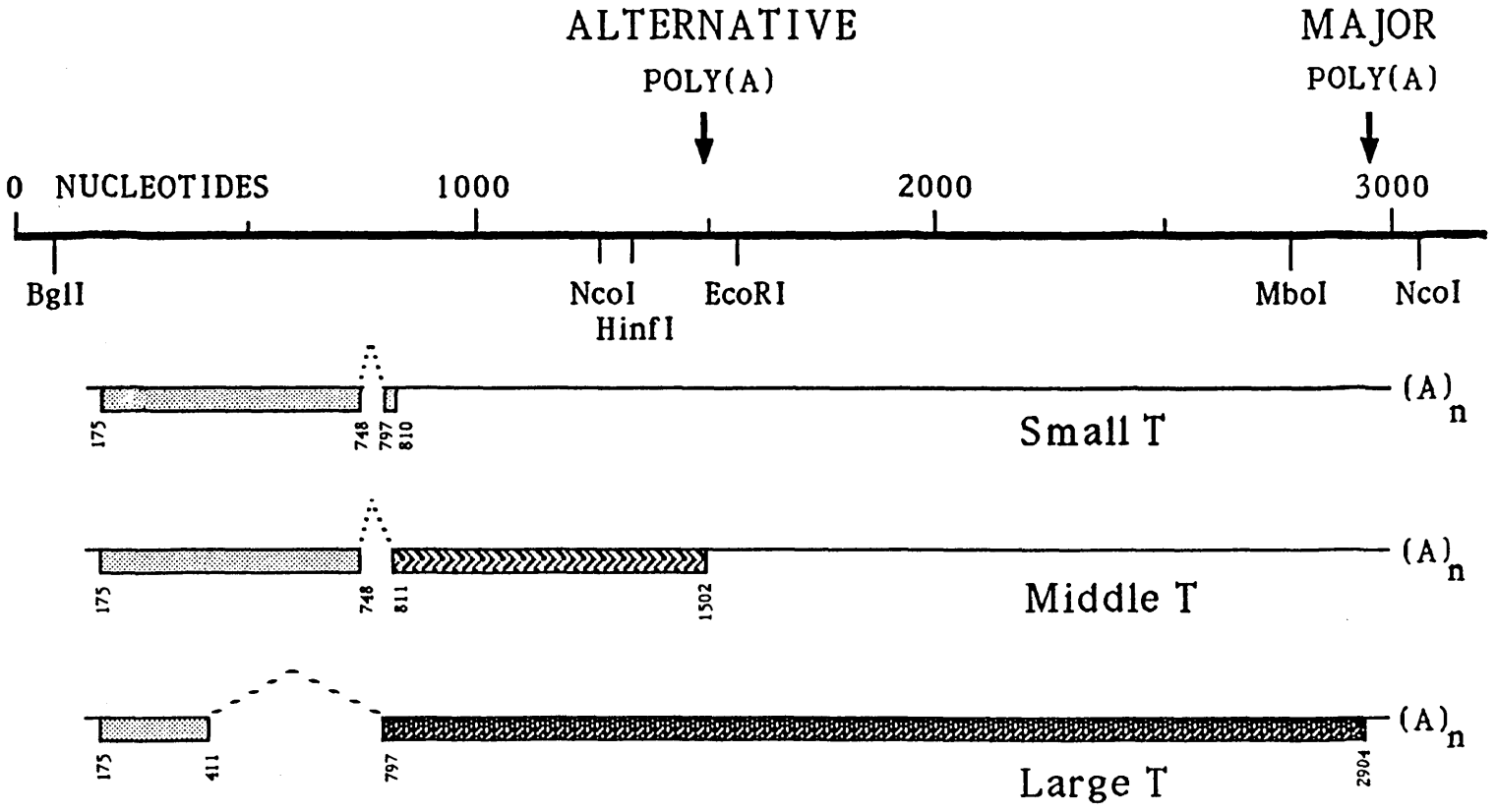
Previous studies of Py early transcription have made use of a variety of techniques to overproduce early mRNAs. Frequently, large T-Ag mutants (tsA) have been used to elicit a stage-specific inactivation of large T protein, with consequent prevention of viral DNA replication and late gene expression accompanied by derepression of early strand transcription. This has proved to be a useful tool, since a major complication of studies of early transcription has been the enormous excess of late strand transcripts which are normally present at late times in infection. Alternative means of favouring early transcription include the use the of DNA replication inhibitors cytosine arabinoside (AraC) and 5-fluorodeoxyuridine (FUdR). It is not clear to what extent mRNAs produced under these conditions, or after tsA "shift-up", retain the structural characteristics of mRNAs produced during the productive cycle *in vivo*. For this reason this study was designed to examine the transcription of wild-type strain Py A2 after infection of mouse 3T6 fibroblasts.

The relative abundances of alternatively polyadenylated and full-length early mRNAs were investigated by northern analysis. S1 nuclease analysis is an alternative technique which could have been used for estimating RNA abundance. Previous investigations of early transcription using the S1 approach have encountered problems in interpretation, largely due to competition of the probe used with the excess of late strand transcripts present at late times in infection. Such competition was observed even when tsA "shift-up" RNA was used (Treisman, 1981). Using northern analysis, the possibility of generating artefactual results through the hybridisation of early RNAs to late transcripts does not arise, since the gel is run under completely denaturing conditions. While the position of the alternative poly(A) site, near the 3' end of the middle T-Ag coding sequences, suggests that the site could be functionally important, the coding capacity of the mRNAs using this site has not previously been described. In order to clarify this

point, the position(s) of cleavage/polyadenylation directed by this alternative signal were investigated, using the complementary techniques of 3' S1 analysis and cDNA cloning. Comparisons were drawn between the alternative and major early poly(A) sites, and 3' S1 analysis was also used to define accurately the position of cleavage/polyadenylation at the latter site. Finally, theoretical considerations were used to explore the extent to which the nucleotide sequence in the region of the alternative poly(A) site is evolutionarily "trapped" by a requirement to conserve the protein sequences encoded in this region.

Figure 2.1.

A linearised representation of the Py early region showing the positions of the known differential RNA processing sites. Nucleotide numbering is adapted from Griffin *et al.*(1981), taking into account the differences in sequence described by Deininger *et al.* (1980). The three differentially spliced predominant early mRNAs encoding small, middle and large T antigens (coding regions boxed) are aligned with the map. The positions of the major and alternative polyadenylation sites are indicated, along with the restriction sites used in this study.


ALTERNATIVE
POLY(A)
MAJOR
POLY(A)
0 NUCLEOTIDES
1000
2000
3000
BglI
NcoI
HinfI
EcoRI
MboI
NcoI
(A)n
175
748
797
810
Small T
(A)n
175
748
811
1502
Middle T
(A)n
175
411
797
2904
Large T

RESULTS

Northern analysis of Py early mRNAs utilizing the major or alternative poly(A) site.

The sizes and amounts of Py early region transcripts using either the alternative or the major poly(A) site (Fig. 2.1) were assessed at different times during the lytic cycle after infection of mouse 3T6 cells with the A2 strain of Py. The relative abundance and overall splicing pattern of early region mRNAs was initially determined by northern blotting (Fig. 2.2). Early region mRNA from the Py transformed cell line 53-Rat, which exclusively utilizes the alternative poly(A) site (Kamen *et al.*, 1979), was analysed in parallel for comparison. The probe used in this analysis was derived from that part of the Py early region common to both alternatively polyadenylated and full-length early mRNAs. Since the extent of homology with the probe was the same for mRNAs of each class, the intensity of the hybridisation signal could be used to estimate the proportion of mRNAs using each site. As the amounts of probe retained on the filter were relatively small, densitometric scanning of the autoradiographs (Fig. 2.3) was used for quantitation in preference to scintillation counting of filter slices. This northern analysis gives rise to two main points of interest. One is the timecourse of appearance of the Py RNA using the alternative poly(A) site; RNA utilizing the alternative site was only detected at relatively late times after infection, accounting for about 15% of the early mRNAs at 22 and 26 hours (Figs 2.2, 2.3). A second point of interest concerns the splicing pattern of the early region mRNAs poladenylated at the two different sites. The ratio of small plus middle T-Ag spliced mRNA to large T-Ag spliced mRNA is 0.5 for the major poly(A) site at all four time points; in contrast, this ratio is 1.3 for the alternative poly(A) site at 22 hours (Fig. 2.3).

Figure 2.2.

Py early region mRNAs expressed during lytic infection of mouse 3T6 cells and in the Py-transformed cell line 53-Rat. (A) 2μg total poly(A)$^+$ RNA from cells harvested at 9, 12, 22 and 26 hours after infection of 3T6 cells with Py A2 or from 53-Rat cells was subjected to 1% agarose/formaldehyde gel electrophoresis. After blotting onto nitrocellulose, the RNAs were probed with a *Bgl*I-*Eco*RI fragment of the Py early region corresponding to nts 87 to 1562 (see Fig. 2.1), which was labelled to high specific activity by random oligonucleotide priming. Early region mRNAs cleaved at the major poly(A) site (MA) and the alternative site (ALT) each appear as a doublet. Within each doublet, the upper band corresponds to the sum of species spliced in the small and middle T-Ag reading frames, whereas the lower band reflects splicing in the large T-Ag reading frame. Py mRNAs in 53-Rat are almost exclusively polyadenylated at the alternative site. Size markers were provided by an RNA ladder (B.R.L.). (B) A shorter exposure of (A), showing the splicing pattern of the major species present at 22 and 26 hours post infection.

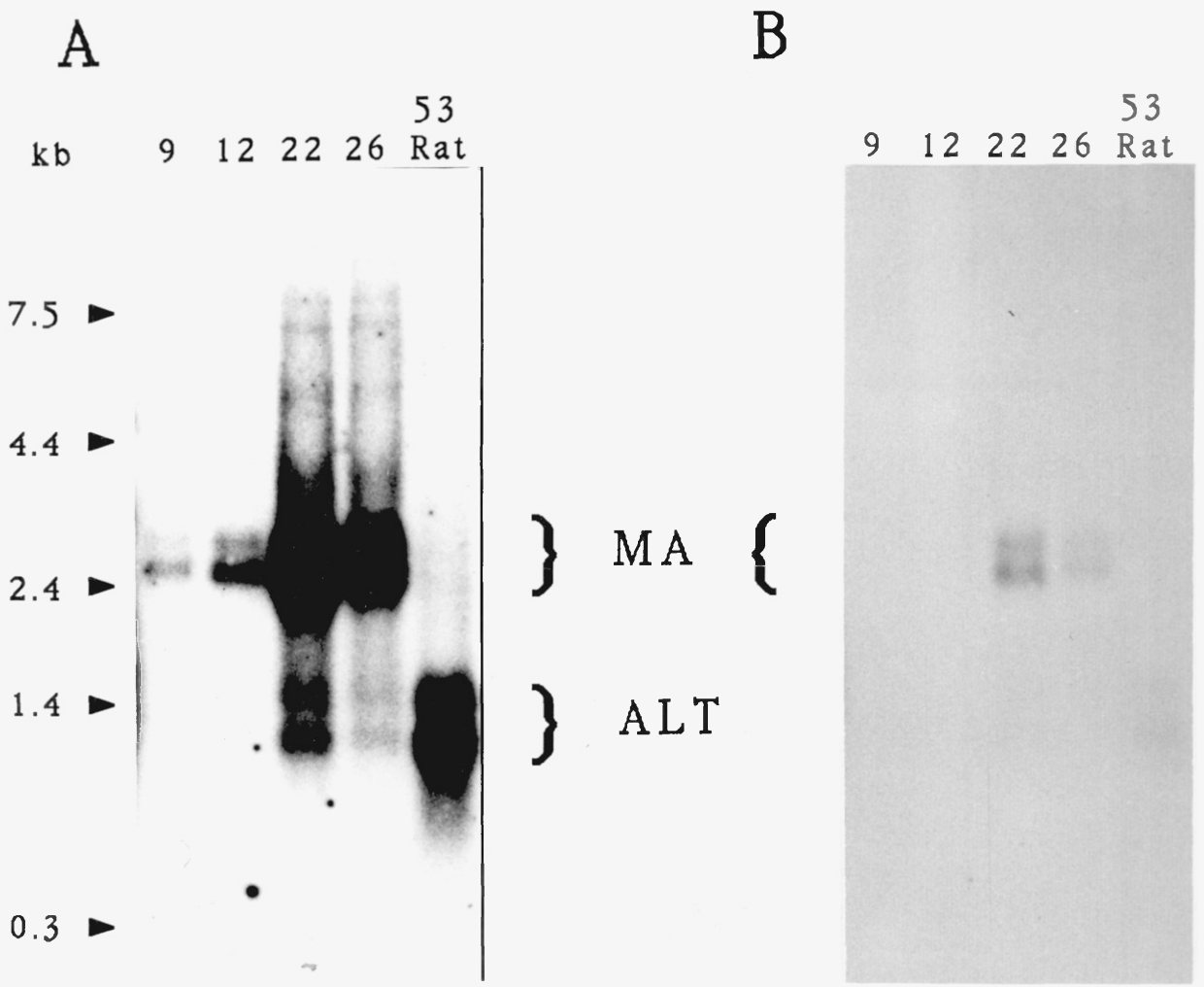
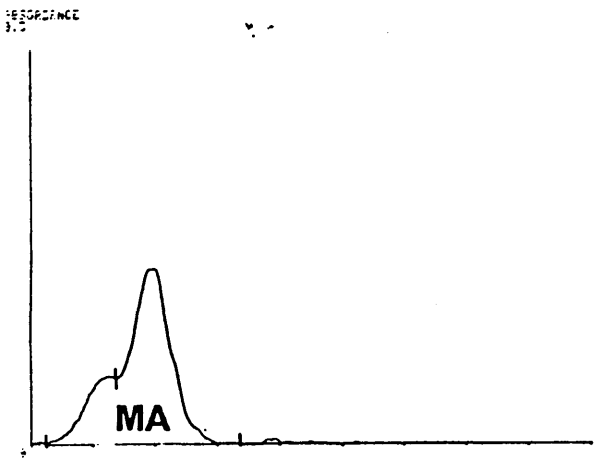
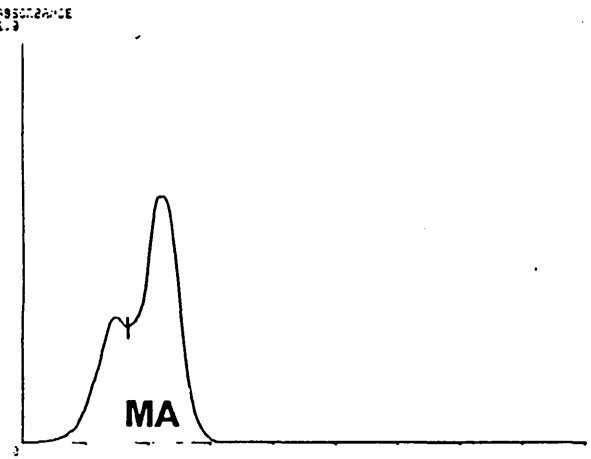

A
B
kb
9
12
22
26
53
Rat
7.5
4.4
2.4
1.4
0.3
MA
ALT

Figure 2.3.

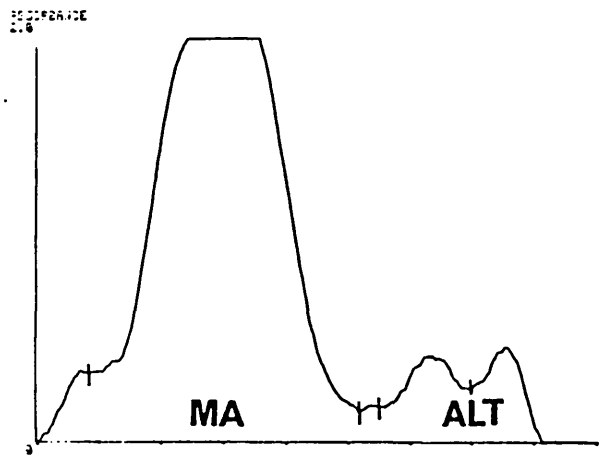
Densitometric scanning of northern blotting data. The autoradiograph shown in Fig 2.2(A) was scanned using a Joyce-Loebl Chromoscan 3 densitometer. Traces obtained by scanning the lanes containing mRNAs from 9, 12, 22 and 26 hours after infection are labelled accordingly. The fifth trace (53) is derived from the lane containing 53-Rat mRNA. Full-scale absorbance was set at 0.5 A for the 9 h lane, 1.0 A for the 12 h lane, and 2.0 A for the remaining lanes. In each case the left-hand side of the trace corresponds to the top of the autoradiograph. The doublets corresponding to major (MA) and alternative (ALT) polyadenylated species are indicated. For each doublet, the left-hand peak represents the sum of species spliced in the small and middle T-Ag reading frames, and the right-hand peak corresponds to mRNAs spliced in the large T-Ag reading frame. For the 22 and 26 h time points, the lighter exposure shown in Fig. 2.2(B) was used to confirm that the splicing ratio associated with the major poly(A) site did not alter during the course of infection.



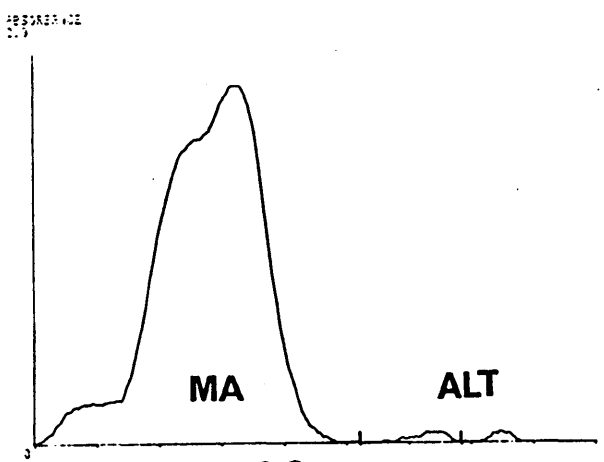
9



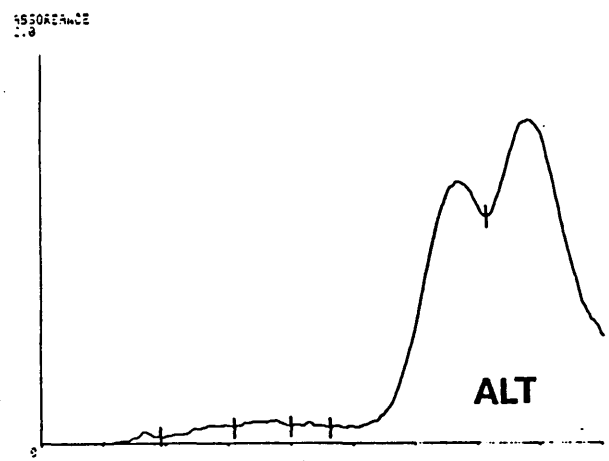
12



22



26



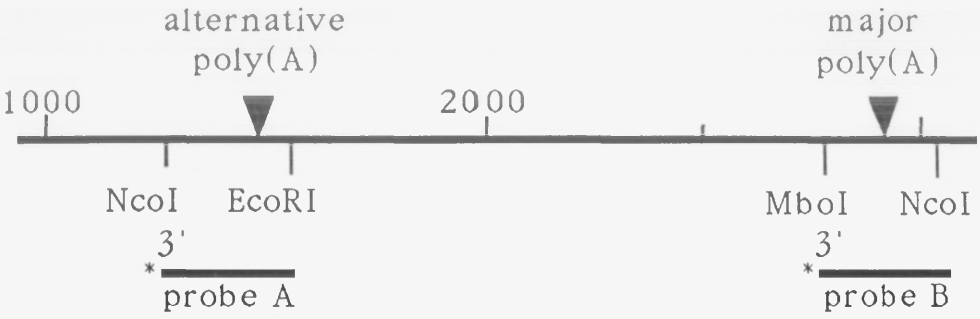
53

3' Heterogenity of early region mRNAs polyadenylated at the alternative site.

The positions of cleavage/polyadenylation at the major and alternative polyadenylation sites were initially investigated by 3' S1 mapping (Fig. 2.4). The 289 nt. *Mbo*I-*Nco*I (Py nts 2771-3060) fragment was protected to give a single band of 171±1 nts, showing polyadenylation of mRNAs at the major site to be at a single position at Py nucleotide 2942±1. This is 17±1 nucleotides downstream from the AATAAA consensus sequence (Fig. 2.6). In contrast to the homogeneity at the major site heterogeneous 3' ends were detected in mRNAs polyadenylated at the alternative poly(A) site (Fig. 2.4). Identical heterogeneous 3' ends were also found for mRNAs from 53-Rat in which the early region transcripts almost exclusively utilize the alternative polyadenylation site. Protection of the 287 nt. *Nco*I-*Eco*RI probe (Py nts 1279-1566) gives three predominant fragments of 216±2, 227±2 and 235±2 nts, mapping the 3' ends of RNAs utilizing the alternative poly(A) signal to three predominant positions (Fig. 2.4). The relative proportions of these bands did not appear to vary between mRNAs harvested at 22 or 26 hours and mRNA from 53-Rat. The sites of end-labelling in the twin-probe experiment (Fig. 2.4) were chosen in order to give well separated protected fragments corresponding to each of the two poly(A) sites on the acrylamide gel system used. As a first step towards verification of the apparent heterogeneity of alternative site polyadenylation, and in order to map more accurately the positions of cleavage, a second S1 experiment was performed (Fig. 2.5). In this experiment the probe was 3' end-labelled at a *Hin*fI site at nt. 1340, closer to the predicted mRNA 3' ends than the *Nco*I-labelled probe described in Fig. 2.4. The generation of S1-resistant fragments of 155, 166 and 174 nts (±1nt.) after annealing with poly(A)$^+$ RNA from 53-Rat serves to substantiate the result shown in Fig. 2.4. Densitometric scanning was used to estimate the relative proportions of RNAs apparently cleaved at the three microheterogeneous sites. About 50% is polyadenylated 12±1 bases downsteam from the AATAAA at Py nucleotide 1495±1, 15% at Py nucleotide 1506±1

Figure 2.4.

3' S1 mapping of Py early mRNAs. End-labelled probes A and B span the alternative and major polyadenylation sites respectively. RNAs (2μg poly(A)$^+$) from PyA2-infected mouse 3T6 cells, harvested at 9, 12, 22 and 26 hours post infection and from 53-Rat were annealed with an equimolar mixture of both probes. A triplet of bands corresponds to alternatively polyadenylated mRNA 3' ends at Py nts 1495±2, 1506±2 and 1514±2. Transcripts utilising the major early poly(A) site give a single band which corresponds to mRNA 3' ends at Py nt. 2942±1 (see text). Note that the major site is absent from the actively transcribed early region in 53-Rat. A negative control (C) was provided by annealing with 40μg of yeast RNA alone. Size markers (M) were end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in the plasmid vector pAT153.


alternative
poly(A)
major
poly(A)
1000
2000
NcoI
EcoRI
MboI
NcoI
3'
3'
probe A
probe B

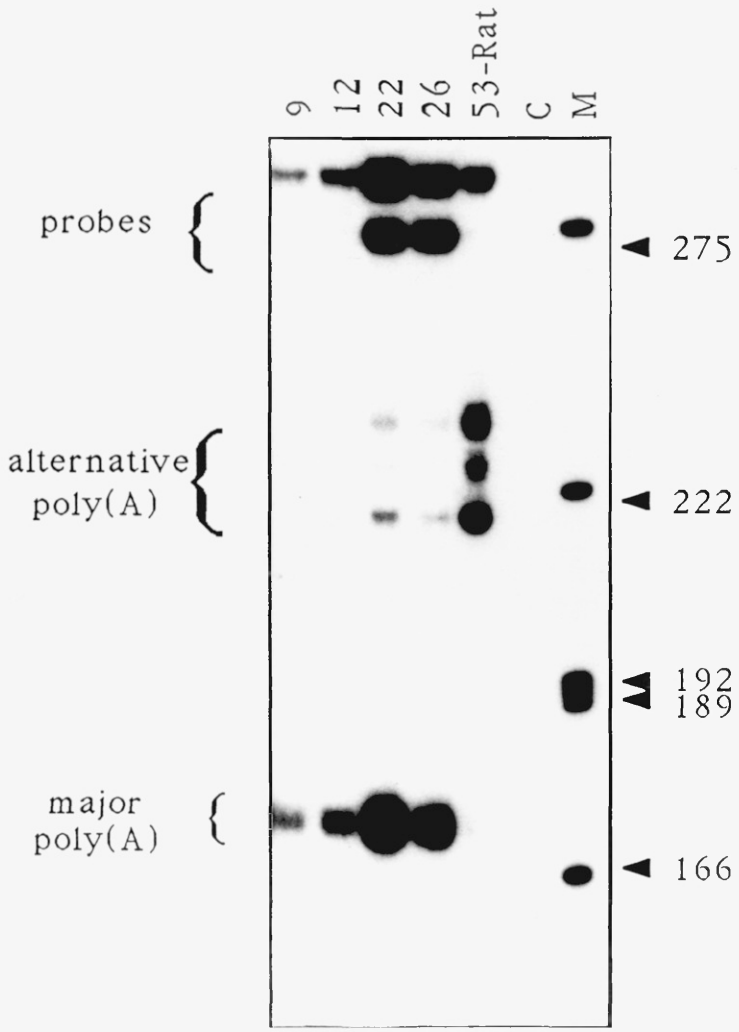
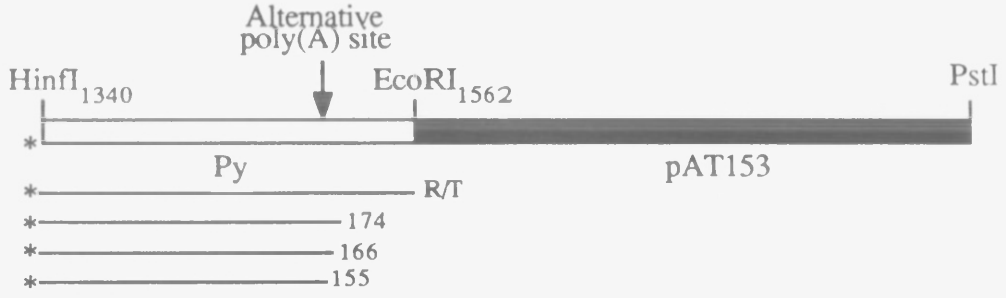
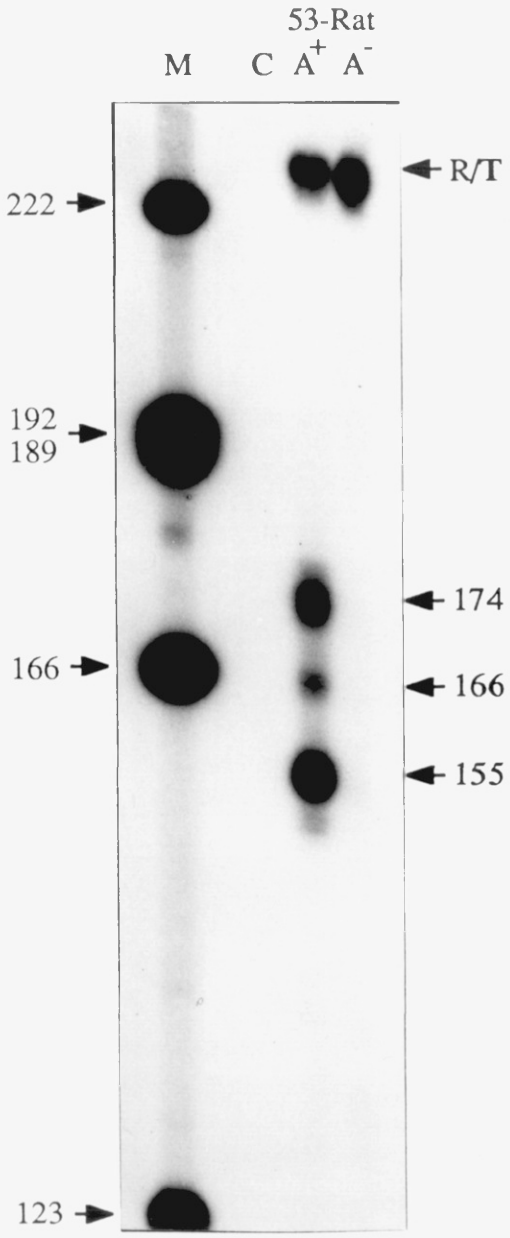

9
12
22
26
53-Rat
C
M
probes
alternative
poly(A)
major
poly(A)
275
222
192
189
166

Figure 2.5.

3' end heterogeneity of RNAs cleaved at the alternative poly(A) site. A probe, represented diagrammatically at the top of the figure, was constructed by 3' end-labelling a *Hin*fI-*Pst*I fragment consisting of Py sequences (white) spanning the alternative poly(A) site fused to plasmid sequences (black) derived from the cloning vector pAT153. The 3' end-label is denoted by an asterisk. The four lines beneath the probe represent the sizes of fragments protected from S1 nuclease digestion after annealing with Py RNAs. Transcripts reading through (R/T) past the viral *Eco*RI site at nucleotide 1562 or cleaved in the region of the alternative poly(A) site protect fragments of the sizes indicated. The lower part of the diagram shows the result of 3' S1 analysis using the probe described above after annealing with 40μg yeast RNA control (C), to 2μg poly(A)$^+$ or to 40μg poly(A)$^-$ RNA from the Py-transformed line 53-Rat. Size markers (M) were end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in pAT153. Estimated sizes of fragments protected by 53-Rat poly(A)$^+$ RNA of 155, 166 and 174 nts (±1 nt.) imply that the alternative signal directs cleavage/polyadenylation at Py nts 1495, 1506 and 1514 (±1 nt.). The densitometric trace of the 53-Rat poly(A)$^+$ lane is shown to the right. The relative intensities of the four signals were 24:26:11:39 (R/T:174:166:155).


Alternative
poly(A) site
HinfI 1340
EcoRI 1562
PstI
Py
pAT153
R/T
174
166
155


53-Rat
M
C
A+
A-
222
192
189
166
123
R/T
174
166
155

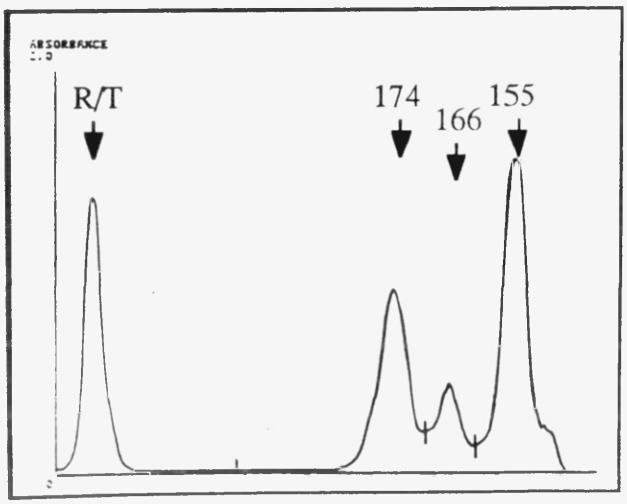
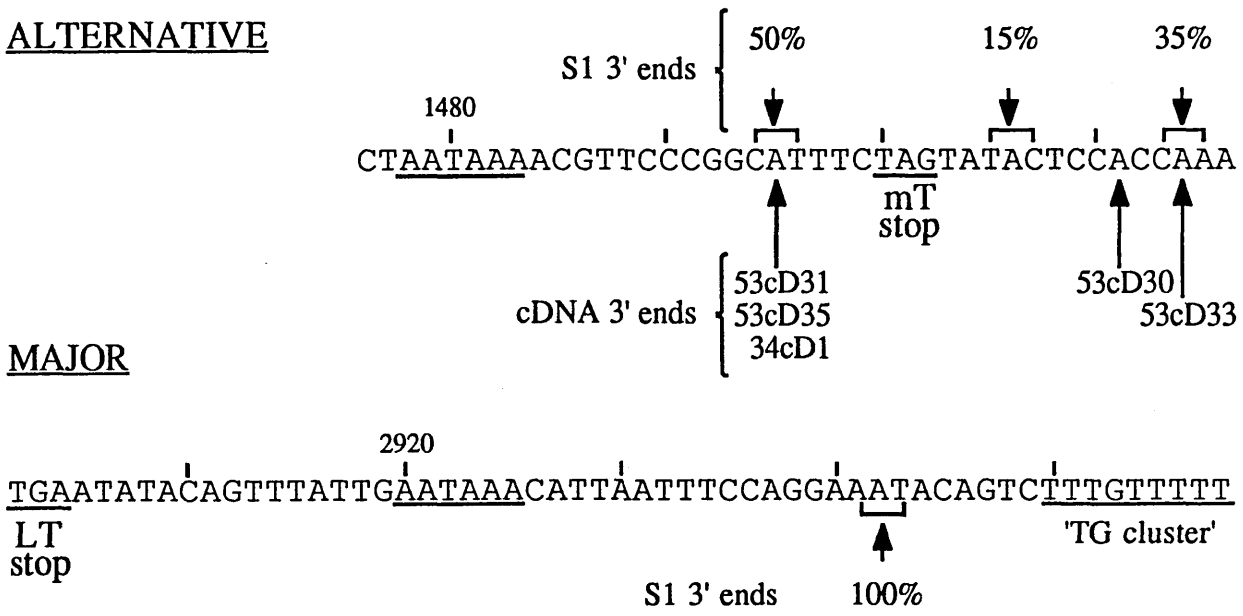

R/T
174
166
155

Figure 2.6.

Positions of cleavage/polyadenylation sites in 53-Rat cDNAs. Oligo(dT)-primed cDNAs were synthesised from 53-Rat total poly(A)$^+$ RNA and were cloned into λGT10. In the upper half of the figure the positions of cleavage/polyadenylation for the five clones sequenced are shown in relation to the AATAAA consensus sequence and the TAG middle T-Ag translational termination codon. The positions and relative abundances of 3' ends determined by densitometric scanning of 3' S1 mapping data are indicated above the sequence. The lower half of the figure shows the deduced position of cleavage/polyadenylation at the major site, in relation to the AATAAA consensus and TGA stop codon in the large T-Ag reading frame.


ALTERNATIVE
S1 3' ends
50%
15%
35%
1480
CTAATAAAACGTTCCCGGCATTTCTAGTATACTCCACCAAA
mT
stop
cDNA 3' ends
53cD31
53cD35
34cD1
53cD30
53cD33
MAJOR
2920
TGAATATACAGTTTATTGAATAAACATTAATTTCCAGGAAATACAGTCTTTGTTTTT
LT
stop
'TG cluster'
S1 3' ends
100%

and 35% at Py nucleotide 1514±1 (see Figs 2.5, 2.6).

This method for mapping sites of polyadenylation may be slightly imprecise since it is known that S1 can digest a few nucleotides past the true end of an RNA-DNA heteroduplex (see, for example, Treisman *et al.*, 1981a). Consequently sequence analysis of the site of polyadenylation of cDNAs derived from early region mRNAs utilising the alternative poly(A) site was undertaken to determine the exact locations of the poly(A) tracts. Since the pattern of alternatively polyadenylated mRNAs in the Py lytic cycle was identical to that observed with 53-Rat, as judged by S1 mapping (Fig. 2.4), this transformed cell line was used as a convenient source of mRNAs for the cDNA cloning. The results of sequence analysis of the points of polyadenylation of five independent cDNA clones are shown in Fig. 2.6 and confirm that, in essence, the S1 result is not artefactual. Three cDNA ends were located at Py nucleotide 1495 and one at Py nucleotide 1514; these are the two major sites identified by S1 mapping. The fifth cDNA end, at Py nucleotide 1511, corresponds to one of the minor heterogenous ends detected by S1 mapping (see Fig. 2.4). Thus both S1 mapping and cDNA analyses show that about 50% of the mRNAs utilizing the alternative poly(A) site are polyadenylated upstream from the middle T-Ag stop codon (Fig. 2.6) and presumably are not capable of directing the translation of a functional middle T-Ag protein.

DISCUSSION

Py early region transcription is complicated not only by differential splicing pathways, but also by the existence of two poly(A) sites. The major site, situated just downstream of the large T-Ag translational stop codon, is used by the majority of all three predominant early region mRNAs (Kamen *et al.*, 1979). An alternative site is signalled by an AATAAA sequence close to the end of the middle T-Ag coding sequences (Fig. 2.1).

Whilst this site is only used at a relatively low level in lytically infected cells in culture, in certain Py-transformed cell lines such as 53-Rat viral transcripts are cleaved and polyadenylated almost exclusively at this position (Kamen *et al.*, 1979). The overall levels of Py mRNAs in these lines are similar to those in lines where the major poly(A) site is used. This suggests that under certain circumstances, the alternative poly(A) site is able to function efficiently, although an alternative explanation is that the rate of transcription of viral sequences is unusually high in lines such as 53-Rat, and that transcription through to a cellular poly(A) site is not possible; in such a case the alternative site would be the only position at which viral-host fusion transcripts could be polyadenylated (see chapter 3). Furthermore, comparison of the Py nucleotide sequence with that of the Hamster Papovavirus, which also appears to encode a middle TAg-like protein (Delmas *et al.*, 1985), shows that the alternative poly(A) site region is very highly conserved. In contrast SV40, which has no middle T-Ag analogue, lacks an alternative poly(A) site, despite the presence of AATAAA consensus sequences within the large T reading frame (Table 1.3).

The major poly(A) site is used by at least 85% of the Py early region mRNAs. About twice as many major site polyadenylated mRNA molecules contain the large T-Ag splice as contain the small and middle T-Ag splices (Figs 2.2, 2.3). The region of the major poly(A) site is similar in its layout to a wide variety of viral and cellular poly(A) sites, with cleavage/polyadenylation occurring at a single site (Fig. 2.4) 17±1 nucleotides (Py nt. 2942±1) downstream from the AATAAA consensus sequence (Proudfoot and Brownlee, 1976) and just upstream from a (T+G)-rich element (McLauchlan *et al.*, 1985; see Fig. 2.6). The alternative poly(A) site has been found to be used by up to 15% of the Py early mRNAs (Fig. 2.2). This maximal level of utilization was observed at later times (22-26 hrs p.i.) in the cycle; at earlier times (9-12 hrs p.i.) alternative polyadenylation was not detected, either by northern (Fig. 2.2) or 3' S1 analysis (Fig2.4), though this may result, in part at least, from the relatively low total signal. The ratio of small and middle T-Ag to

large T-Ag splices for the alternatively polyadenylated RNAs is about 2.5 times that observed for mRNAs using the major site. The splicing pattern associated with the alternative site would thus favour the production of small and/or middle T-Ag, with less of the large T-Ag splice, which would not give rise to a known protein product, being produced. As the alternative poly(A) site occurs in the middle of the large T-Ag coding region (Fig. 2.1) no pre-existing translational termination codon exists in the large T-Ag reading frame before the end of the large T-Ag spliced mRNA. This is consistent with the fact that no large T-Ag related proteins have been reported from the translation of the alternatively spliced large T-Ag mRNAs, either *in vivo* or *in vitro*, from the lytic cycle or from 53-Rat (Kamen *et al.*, 1979; Lania *et al.*, 1980).

The sites of polyadenylation governed by the alternative signal are heterogenous (Fig. 2.4), occurring at at least three distinct sites between Py nucleotides 1495 and 1514 (Fig. 2.6). There are some minor discrepancies between the results of the S1 and cDNA analyses. Thus, while four of the five cDNA 3' ends are in accord with the 3' S1 analyses, one (53cD30) does not correlate with an S1-resistant product (see Fig. 2.6). This difference probably serves to underline the degree of inaccuracy inherent in the S1 technique, though the possibility remains that 53cD30 was generated by an aberrant oligo(dT) priming event. Heterogenity with respect to the site of polyadenylation governed by a single poly(A) signal is unusual, but has been described for three other transcription units, namely the bovine prolactin, mouse ribosomal protein L30 and hepatitis B surface antigen genes (Simonsen and Levinson, 1983; Sasavage *et al.*, 1982; Wiedeman and Perry, 1984). Comparison of the nucleotide sequences of these sites with that of the alternative Py early site does not reveal any obviously common features which might account for the observed cleavage heterogeneity. In particular, although the Py alternative site is the only functional Py or SV40 poly(A) site to lack the UGU motif (Table 1.3), this motif is present, between 15 and 31 nts downstream from the AAUAAA consensus, in all three of the other heterogeneously polyadenylated transcripts (Simonsen and Levinson,

1983; Sasavage *et al.*, 1982; Wiedeman and Perry, 1984).

Whereas the major poly(A) signal is located downstream from the large T-Ag reading frame stop codon at the end of the Py early region the alternative poly(A) signal is located in the middle of the early region upstream from the stop codon of the middle T-Ag reading frame (see Fig. 2.6). Thus none of the alternatively polyadenylated mRNAs could produce a functional large T-Ag; it would not be expected that a truncated form of large T-Ag would be efficiently translated as no pre-existing stop codon exists in the large T-Ag reading frame before the ends of the alternatively polyadenylated RNAs. Interestingly polyadenylation at Py nucleotide 1506 would result in the creation of a TAA stop codon in the large T-Ag reading frame (Fig. 2.6) and such an mRNA could code for a truncated large T-Ag of 35K. It would be expected that such a protein might be present only in small amounts, as S1 mapping suggests that only 15% of the alternatively polyadenylated mRNAs are polyadenylated at Py nt. 1506 and this site was not detected by the cDNA analysis (Fig. 2.6). In addition the alternatively polyadenylated mRNAs contain relatively little of the large T-Ag splice (Figs 2.2, 2.3).

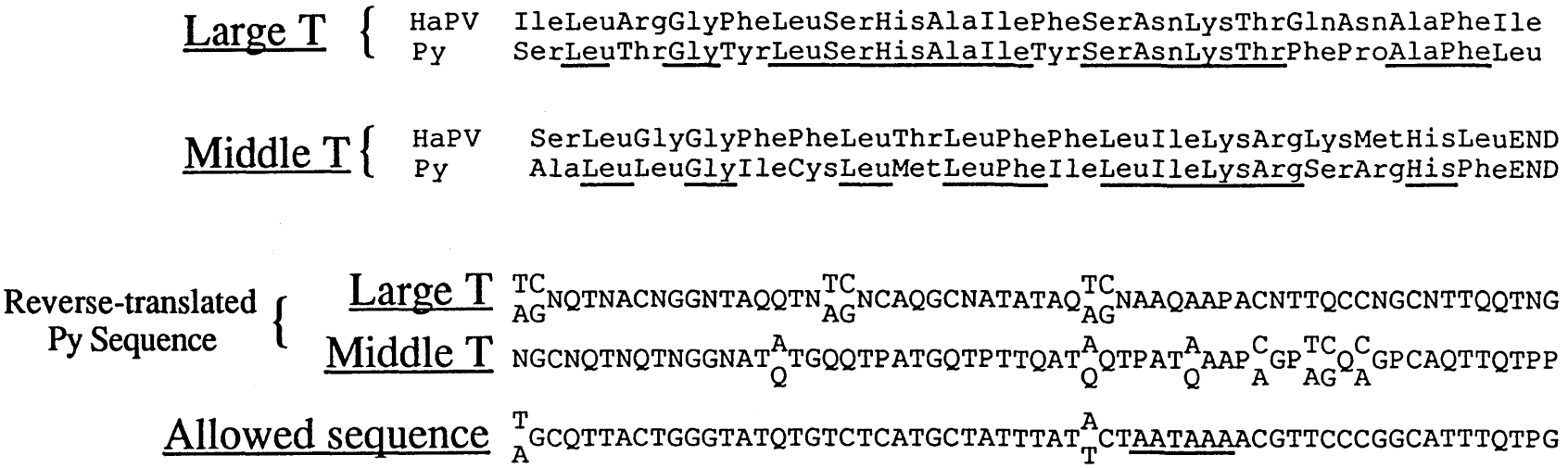
The biological consequences of these results are that whilst all alternatively polyadenylated species could, if appropriately spliced, direct the synthesis of small T-Ag, only the 50% polyadenylated downstream from the middle T-Ag stop codon could encode middle T-Ag. Transcripts cleaved at the alternative poly(A) site and spliced in the large T-Ag reading frame do not appear to give rise to a truncated large T-Ag-related protein; by analogy, transcripts spliced in the middle T-Ag frame and polyadenylated at nt. 1495 would not be expected to direct the synthesis of a truncated middle T-Ag. It will be interesting to discover whether or not cleavage at one or other of the alternative positions is associated with a particular combination of splice sites. A correlation between selection of a particular poly(A) site and selection of a combination of alternative splice sites has been demonstrated in the case of the adenovirus major late transcription unit (reviewed by Nevins, 1983). Further evidence of a cause-effect relationship between alternative

polyadenylation and differential splicing has come from studies of the immunoglobulin heavy chain, calcitonin, GAR transformylase and tropomyosin genes (Alt *et al.*, 1980; Early *et al.*, 1980; Amara *et al.*, 1982; Henikoff *et al.*, 1983; Helfman *et al.*, 1986; see chapter 1).

Although no function has yet been ascribed to middle T-Ag through studies of the lytic cycle in *in vitro* tissue culture systems, this does not appear to be the case for viral replication *in vivo* in the animal. Sequences spliced out of the large T-Ag mRNA but retained in the mRNAs for small and middle T-Ag have been shown to be essential for viral growth in the mouse (McCance, 1981). Indeed, the virus would not be expected to conserve the middle T-Ag reading frame if the protein had no function in its life cycle in the mouse. Comparison of the nucleotide sequences of Py and HaPV has shown that the region of the Py alternative early poly(A) site is conserved in the hamster virus (Delmas *et al.*, 1985; see Fig. 2.7). A possible explanation for the high degree of DNA sequence homology between the alternative poly(A) site of Py and the corresponding region of HaPV is that the AATAAA sequence could encode important amino acids of both the middle and large T-Ag proteins, in overlapping reading frames. Certainly, four amino acid residues (Ser-Asn-Lys-Thr) in the large T-Ag frame and four (Leu-Ile-Lys-Arg) in the middle T-Ag frame are encoded by the DNA sequence spanning the alternative poly(A) hexanucleotide (Fig. 2.7) and are conserved between the Py and HaPV proteins. This situation may constitute an evolutionary trap, in that the possibilities for divergence at the level of DNA sequence could be very limited by a requirement to conserve the two protein sequences. A much more attractive hypothesis would be that small and middle T-Ag synthesis might be regulated *in vivo* by control of polyadenylation at the alternative site. One possibility consistent with the data presented here is that the alternative site could be brought into play late in Py infection in order to increase the abundance of mRNAs encoding middle, and more especially small T-Ag. A mounting body of evidence suggests that one or both of these T-Ags is required at late times in infection to perform a function

Figure 2.7.

Protein sequences encoded in the region of the Py and HaPV alternative poly(A) sites. The upper half of the figure shows the protein sequences predicted for Py large and middle T-Ags by translation, in the two overlapping reading frames, of the DNA sequence in the region. These protein sequences are aligned with those similarly derived from the hamster papovavirus (HaPV) genomic sequence. Predicted amino acid residues held in common between the two viral sequences are underlined. The lower half of the figure shows the "reverse-translated" sequences for the Py polypeptides aligned with the protein sequences from which they were derived. Positions at which any nucleotide (N), pyrimidine residues (Q) and purine residues (P) could be substituted without alteration of each protein product are indicated. The bottom line represents a summation of the two reverse-translated sequences, and is the only sequence which can encode, in overlapping reading frames, the two Py polypeptides shown. The alternative Py polyadenylation site hexanucleotide consensus is underlined.


Large T
HaPV IleLeuArgGlyPheLeuSerHisAlaIlePheSerAsnLysThrGlnAsnAlaPheIle
Py SerLeuThrGlyTyrLeuSerHisAlaIleTyrSerAsnLysThrPheProAlaPheLeu
Middle T
HaPV SerLeuGlyGlyPhePheLeuThrLeuPhePheLeuIleLysArgLysMetHisLeuEND
Py AlaLeuLeuGlyIleCysLeuMetLeuPheIleLeuIleLysArgSerArgHisPheEND
Reverse-translated Py Sequence
Large T
TC/AG NQTNACNGGNTAQQTN TC/AG NCAQGCNATATAQ TC/AG NAAQAAPACNTTQCCNGCNTTQQTNG
Middle T
NGCNQTNQTNGGNAT A/Q TGQQTPATGQTPTTQAT A/Q QTPAT A/Q AAP C/A GP TC/AG Q C/A GPCAQTTQTPP
Allowed sequence
T/A GCQTTACTGGGTATQTGTCTCATGCTATTTAT A/T CTAATAAAACGTTCCCGGCATTTQTPG

in viral assembly, possibly related to modification of the viral capsid proteins (Garcea and Benjamin, 1983). One advantage of using small and middle T-Ag mRNAs cleaved at the alternative, rather than the major, early poly(A) site would be that the former (but not the latter) mRNAs would be free from antisense inhibition (Harland and Weintraub, 1985) by the late mRNAs present at high levels at late times in infection. *In vitro* mutagenesis could not be used to destroy the alternative poly(A) site AATAAA hexanucleotide without altering the protein sequence of either large or middle T-Ag (Fig. 2.7). There are, however, a number of possible point mutations which would be predicted to impair the alternative poly(A) site function while altering the protein sequence of only one of the two T-Ags. For example, mutation of the AATAAA hexanucleotide to AACAAA should drastically reduce the efficiency of polyadenylation at the alternative site (Wickens and Stephenson, 1984). This base change would not alter the large T-Ag protein sequence, but would cause an Ile to Thr change near the middle T-Ag C-terminus. In order to distinguish phenotypic effects due to impairment of the alternative poly(A) signal from those effects which might be due to alteration of protein sequence, the wild-type version of the altered protein could be supplied in *trans* from a second mutant virus. For the example given above, the second virus could contain a mutation of the AATAAA sequence to GATAAA; this sequence is not known to function in signalling polyadenylation, but the viral mutant would encode a wild-type middle T-Ag.

CHAPTER 3: SEQUENCES DOWNSTREAM FROM THE VIRAL INSERT IN 53-RAT DIRECT PREFERENTIAL USE OF THE POLYOMA VIRUS ALTERNATIVE POLYADENYLATION SITE

53-Rat is one of a small number of Py-transformed cell lines in which the Py alternative early poly(A) site is known to be used at a relatively high level. Using cloned sequences located downstream from the viral insert in this line, the extent to which these sequences are responsible for the unusual poly(A) site preference was investigated. In plasmid-based transient expression assays, the 53-Rat downstream sequences were tested for their capacity, when placed downstream from the Py early promoter and between the alternative and major poly(A) sites, to prevent transcripts from using the efficient downstream poly(A) site. By using a 1.4 kb fragment of 53-Rat downstream cellular DNA in the assays, it was possible to reproduce the polyadenylation phenotype found in the parental cell line; readthrough transcription to the major poly(A) site was largely prevented by the interposition of this fragment. Attempts to define a smaller sub-fragment containing the full transcriptional interference activity were unsuccessful; this may indicate that the element(s) responsible are complex. The sequence of the 1.4 kb fragment revealed some similarities to previously described POLII terminators, but poly(A) site-independent termination activity could not be demonstrated for the 53-Rat sequence. A surprising result was that a novel Py 5' splice site was found to be activated by the presence downstream of a compatible rat 3' splice site. Both sites were identified and their sequences were compared to those of previously described Py splice sites.

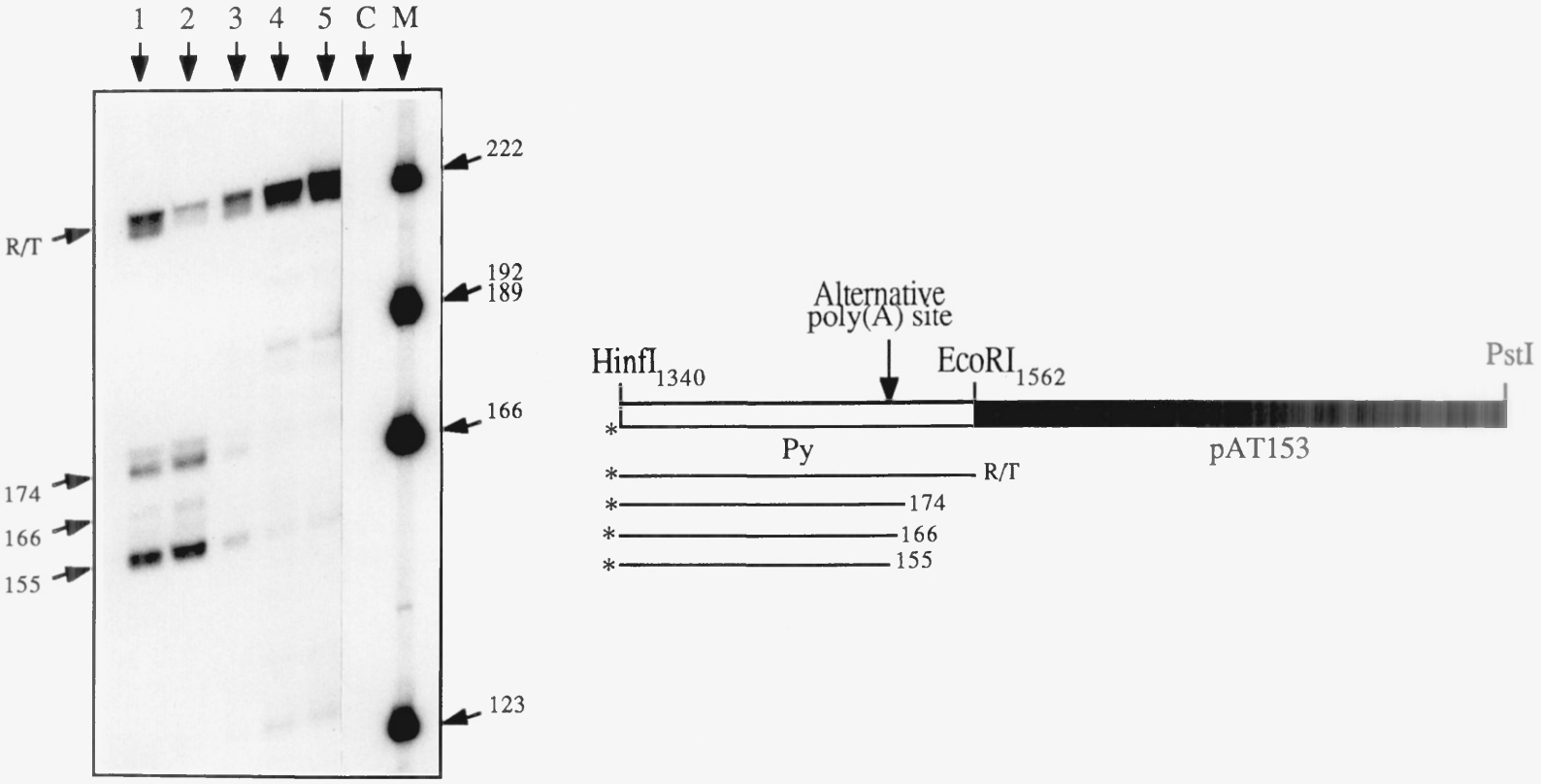
OBJECTIVES

The selection against maintenance of an intact large T-Ag gene which appears to occur in Py-transformed cells (see chapter 1) can have significant consequences for Py transcript processing. In cases where the large T-Ag sequences are disrupted by point mutation or by insertion/deletion, all early region splice sites and both poly(A) sites can be retained. In such transformed lines the pattern of processing of viral transcripts has been shown to resemble that found at early times in the Py lytic cycle, with the majority of transcripts being cleaved at the major early poly(A) site (Hayday *et al.*, 1982; Ruley and Fried, 1983); this might be expected, as in each case the template contains identical regulatory signals, and use of these signals is not likely to be complicated by antisense transcription from the late promoter. Such retention of the full set of early transcriptional signals is the exception rather than the rule, as judged from the characterisation of a number of Py transformed lines (Fried *et al.*, 1986: see appendix 2). More typically, recombination between viral and host sequences, either during or after the initial integrative event, generates a truncated early region lacking C-terminal large T-Ag sequences (Fried *et al.*, 1986: see appendix 2). The major early poly(A) site is consequently lost, with retention of the alternative site being brought about by concurrent selection for maintenance of the middle T-Ag. How, then, is a truncated early region to achieve efficient middle T-Ag expression, lacking as it does its major poly(A) site? Where singly integrated, truncated Py early regions have been studied, following selection for a transformed phenotype, efficient polyadenylation in downstream cellular sequences is often found to be the means by which the required level of middle T-Ag expression is achieved (Fenton and Basilico, 1981; Ruley *et al.*, 1982; B. Davies, personal communication). These results are in agreement with those of Santangelo and Cole (1983), which showed that a functional poly(A) site is found, on average, every 4 kb in randomly isolated genomic DNA fragments (see chapter 1).

The least common means by which efficient polyadenylation of viral mRNAs has been shown to be achieved in Py-transformed cells is exemplified in the lines tsc3T3-2 and 53-Rat (Kamen *et al.*, 1979). In both lines the transcriptionally active early region lacks C-terminal large T-Ag sequences and their associated major poly(A) site. Just as in other transformed lines containing 3'-truncated early regions, hybrid nuclear transcripts initiated at the viral early cap sites extend into the cellular sequences downstream. In the majority of transformed lines these hybrid primary transcripts would be cleaved and polyadenylated at cellular sites fortuitously located nearby. In contrast, the hybrid nuclear transcripts of tsC3T3-2 and 53-Rat are cleaved and polyadenylated almost exclusively at the alternative early poly(A) site (Kamen *et al.*, 1979). Furthermore, the amounts of Py mRNA produced in these lines are of the same order as those found in lines wherein the Py major early or fortuitous cellular poly(A) sites are used (Kamen *et al.*, 1979 this study, Fig. 3.1). On the face of it, it would appear that the alternative site is used more efficiently in tsC3T3-2 and 53-Rat than it is either in the productive cycle or in the majority of Py-transformed lines. If an efficient cellular poly(A) site is present downstream from the viral inserts in these lines, and the results of Santangelo and Cole (1983) suggest that this is likely, then the alternative Py site is chosen preferentially as the site of cleavage. Even if there is no cellular poly(A) site nearby, why should such a high level of Py nuclear RNAs be cleaved at the alternative site, which accounts for at most 15% of the lytic cycle early 3' ends (see chapter 2)? It is possible that, if the proportion of nuclear transcripts cleaved at the alternative site is fairly constant in different Py transformed lines, and if there is a very high level of nuclear transcription in lines such as tsC3T3-2 and 53-Rat, this could account for the correspondingly high level of alternatively polyadenylated mRNAs. If cleavage at the alternative site is an inefficient process, a large reserve of nuclear Py transcripts would be required in order to maintain the relatively high level of Py mRNA present in these lines. The observation that only low levels of Py nuclear transcripts are present in tsC3T3-2 and 53-Rat cells (Treisman, 1981; this study) therefore lends weight

Figure 3.1.

A survey of alternative poly(A) site use in Py-transformed rodent cell lines. 3' S1 nuclease analysis was performed using the Py/pAT153 *Hin*fI-*Pst*I probe described in Fig. 2.5 and shown here on the right-hand side, and 40μg total cell RNA from Py-transformed lines 53-Rat (lane 1), tsc3T3-2 (lane 2), wt3T3-1 (lane 3), 7axB (lane 4), 7axT (lane 5), or 40μg yeast RNA control (C). In each case RNAs reading through past the viral *Eco*RI site protect a band of about 222 nts (R/T), whereas RNAs cleaved at the alternative poly(A) site give rise to fragments in the range 155-174 nts.


1 2 3 4 5 C M
R/T
174
166
155
222
192
189
166
123
Alternative
poly(A) site
HinfI1340
EcoRI1562
PstI
Py
pAT153
R/T
174
166
155

to the argument that it is the frequency with which nuclear transcripts are cleaved at the alternative site, rather than the absolute rate of nuclear transcription, which determines the high level of alternative polyadenylation in these lines. Since the 5' expression sequences present in each Py-transformed line are identical, one must conclude that the differences in 3' processing reflect the different cellular locations in which each viral transcription unit has integrated. With this in mind, experimental strategies were designed to investigate the host sequences which might be responsible for enhanced alternative polyadenylation in 53-Rat cells. It was thought possible, at the outset, that it might be the absence of a downstream cellular poly(A) site in this line which causes an unusually high proportion of nuclear Py transcripts to use the Py alternative site. No downstream site would be present if, for example, the viral integrant in this line is adjacent to a chromosomal breakpoint or telomere. Alternatively, the 3' flanking cellular sequences might act to prevent the Py hybrid transcripts from using a downstream, efficient poly(A) site. These possibilities were tested by insertion of a known, efficient poly(A) site, the Py early major site, downstream of the flanking cellular sequences in plasmid reconstructions of the 53-Rat active Py transcription unit. In transient expression studies using these plasmids, it was possible to investigate the fate of hybrid nuclear Py transcripts derived from the modified 53-Rat transcription units. Sub-fragments of the cloned region flanking the Py insert in 53-Rat were used in an attempt to define sequence element(s) responsible for the unusual polyadenylation pattern observed in this line. The possibility that the 53-Rat flanking sequences act as a self-contained POLII terminator was tested by inserting these sequences upstream from the poly(A) site of a heterologous test plasmid, pSV2CAT, which contains the bacterial chloramphenicol acetyl transferase (CAT) gene.

RESULTS

Use of the Py alternative early poly(A) site in transformed lines.

3' S1 nuclease analysis was used to assess the extent to which the alternative site is used in Py RNA polyadenylation in five transformed lines (Fig. 3.1). The experiment was designed in such a way that RNAs extending beyond the viral *Eco*RI site all contribute to a single protected fragment (~222 nts), whereas transcripts cleaved at the alternative Py poly(A) site protect fragments in the range 155-174 nts. Comparison of band intensities for the two sets of protected fragments within each lane thus gives a measure of the extent to which the alternative viral site is used, relative to downstream viral or cellular poly(A) sites. Since equal amounts of RNA were used for each annealing, comparison of the total signal in each lane gives an indication of the relative steady-state levels of Py RNAs in these lines. Of the lines examined, only 53-Rat and tsC3T3-2 displayed an appreciable level of cleavage at the alternative early poly(A) site. wt3T3-1 and 7axT transcripts are polyadenylated predominantly at cellular poly(A) sites downstream from the viral inserts (Treisman, 1981; Wilson, 1986); 7axB contains a partial tandem triplication of Py sequences, and the majority of Py transcripts in this line utilise the Py major early poly(A) site (Hayday, 1982). The approximately equal levels of total signal in each lane suggest that overall levels of Py RNA are similar in all five transformed lines, regardless of their poly(A) site preference.

Viral-host fusion transcripts in 53-Rat.

Earlier studies (Kamen *et al.*, 1979; Treisman, 1981) have shown that nuclear Py transcripts in 53-Rat cells extend at least as far as the viral-host junction at nt. 2314 (Fig.

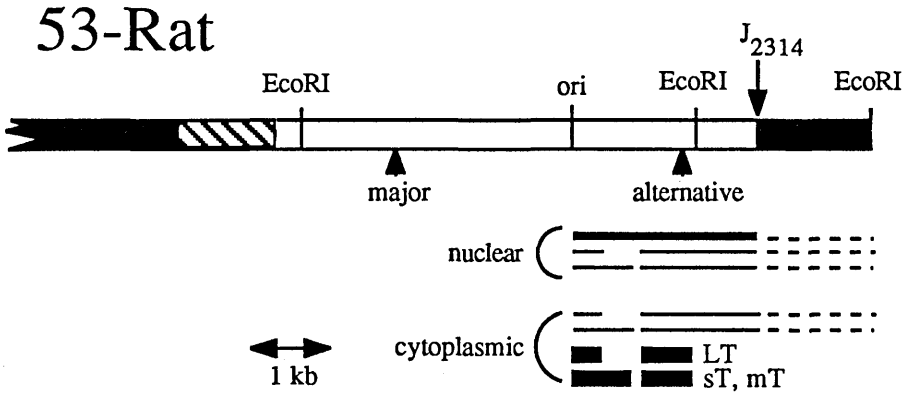


Figure 3.2.

A transcriptional map of the viral sequences in the Py-transformed line 53-Rat (adapted from Treisman, 1981). The viral insert in this line is present as a partial tandem duplication, with the Py sequences (white) bounded by rat DNA (black). The hatched region indicates that the precise position of the left-hand junction is unknown. The right-hand viral-cell junction is at Py nt. 2314 (Ruley *et al.*, 1987) and the right-hand *Eco*RI site of the genomic map marks the 3' extent of sequences cloned from this line. The left-hand partial early region lacks 5' transcriptional control sequences, and transcription is consequently only detected from the 3' truncated, right-hand early region. Nuclear and cytoplasmic RNA species are aligned with the DNA map. In S1 nuclease analyses using Py probes (Treisman, 1981), it has been found that nuclear Py transcripts in this line extend as far as, and presumably beyond, the right-hand viral-host junction. These primary transcripts are cleaved and polyadenylated almost exclusively at the alternative poly(A) site as indicated; the abundance of each RNA species is indicated by line thickness.

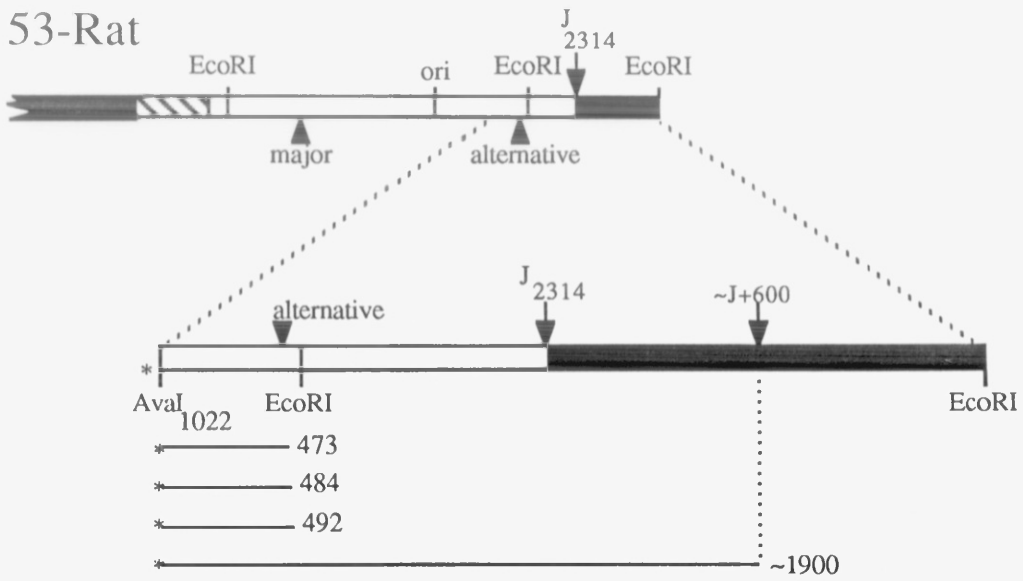
3.2). Prior to the isolation of cloned rat sequences originating from the region to the right of the 53-Rat viral insert (as drawn in Fig. 3.2), it was not possible to investigate the structure of these viral-host fusion transcripts more fully. A λgt-WES clone containing the right-hand *Eco*RI fragment which spans the viral-host junction at Py nt. 2314 has now been isolated (Ruley et al., 1987); availability of this clone, termed 53J, has made possible further investigation of the structure of viral-host fusion transcripts in this line.

A 3' S1 experiment was carried out using a probe labelled at the *Ava*I site at nt. 1022, upstream from the alternative poly(A) site, and extending through to the downstream rat *Eco*RI site (Fig. 3.3). This probe detected the expected RNA 3' ends in the vicinity of the alternative poly(A) site, giving protected fragments in the range 470 to 490 nts; these fragments accounted for the majority of the signal detected using 53-Rat total cell RNA. A fragment of ~1900 nts was also protected by this probe, indicating the presence, at a relatively low level, of apparent 3' ends some 600 nts downstream from the viral-host junction. The fact that trace amounts of full-length probe survived S1 digestion after annealing with 53-Rat total RNA suggests that a correspondingly small proportion of the steady-state RNA may extend beyond the promoter-distal *Eco*RI site. The status and position of the low-level 3' ends mapping within the downstream cellular sequences was investigated further using a 3' end-labelled probe specific for this region (Fig. 3.4). This probe, labelled at an *Nco*I site 340 nts downstream from the viral-host junction, detected apparent 3' ends in both poly(A)$^+$ and poly(A)$^-$ RNAs from 53-Rat. A band common to both RNA fractions of 116 nts (±1nt.) indicated the presence of 3' ends in the region of 456 nts downstream from the viral-host junction. In addition a 230 nt. (±2 nts) band was protected using poly(A)$^+$ RNA. In experiments using the same probe with greater S1 nuclease concentrations the relative yield of the 116 nt. band was increased, suggesting that this position might correspond to an S1-hypersensitive site in the RNA-DNA duplex. These data are discussed in the context of the local DNA sequence in a later section.

The high frequency of cleavage at the Py alternative poly(A) site observed in 53-Rat

Figure 3.3.

Apparent 3' ends of Py transcripts in 53-Rat total cellular RNA. The 3' end-labelled probe, constructed from plasmid p220.2 (see Fig. 3.5) and represented in the upper half of this figure, is contiguous with 53-Rat sequences from the *Ava*I site used for labelling to the *Eco*RI site at the right-hand extent of the cloned sequences. S1-resistant products obtained after annealing the probe with 40μg total RNA from 53-Rat (lane 1) or to 40μg yeast RNA (lane 2) are shown in the lower half of the figure. The signal obtained by omitting S1 nuclease from the reaction after annealing with yeast RNA is shown in lane 3 (in lighter exposures this signal appeared as a single band). The estimated sizes of bands corresponding to alternative site polyadenylation are between approximately 470 and 490 nts; the sizes of these fragments predicted by the results described in Fig. 2.6 are shown beneath the probe diagram. An additional low-level protected fragment of approximately 1900 nts maps an apparent RNA 3' end to about 600 nts downstream from the viral-host junction at Py nt. 2314. Size markers were end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in pAT153.


53-Rat
EcoRI
ori
EcoRI
J
2314
EcoRI
major
alternative
alternative
J
2314
~J+600
*
AvaI
1022
EcoRI
EcoRI
473
484
492
~1900

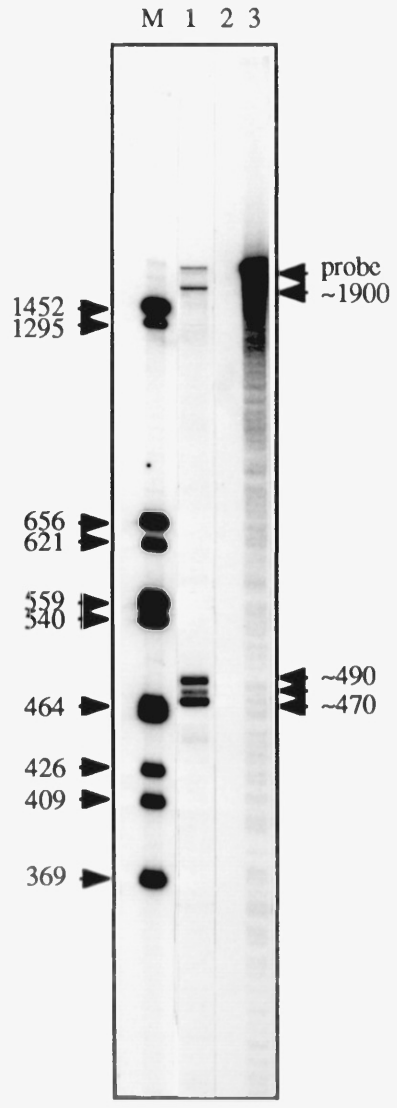
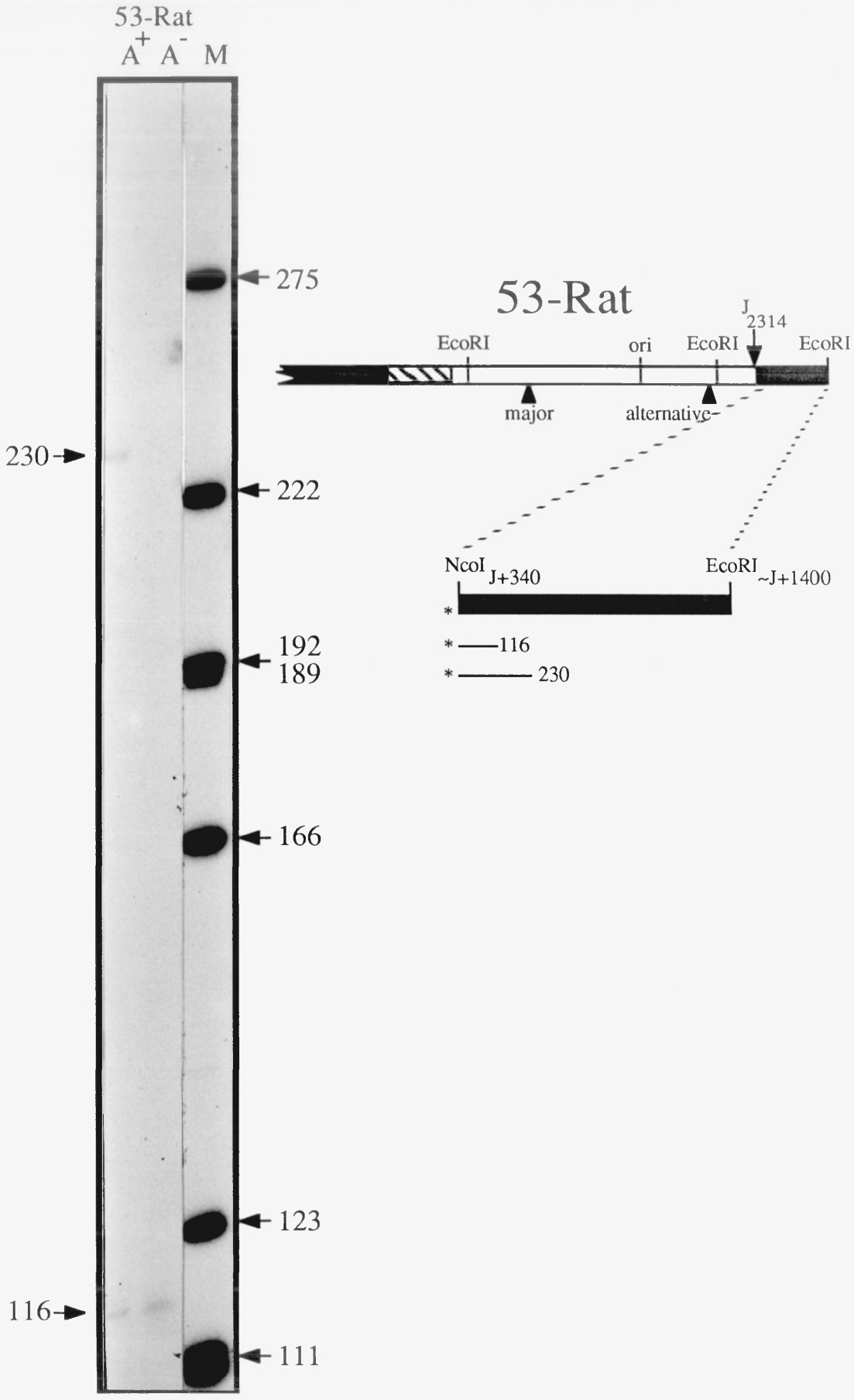

M 1 2 3
probe
~1900
1452
1295
656
621
559
540
464
426
409
369
~490
~470

Figure 3.4.

Accurate mapping of apparent 3' ends of viral-host fusion transcripts in 53-Rat. The probe for this experiment was 3' end-labelled at the *Nco*I site 340 nts downstream from the viral-host junction (J+340) and extended as far as the right-hand *Eco*RI site, approximately 1400 nts downstream from the junction (J+1400). The probe was protected from S1 nuclease digestion by both poly(A)$^+$ and poly(A)$^-$ RNAs from 53-Rat (2 and 40 μg respectively) to give a fragment of 116 nts (±1 nt.). In addition, a fragment of 230 nts (±2 nts) was protected by the poly(A)$^+$ fraction. Representations of these fragments are aligned with the probe diagram, with asterisks denoting the site of labelling. Size markers were end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in pAT153.


53-Rat
A+ A- M
275
230
222
192
189
166
123
116
111
53-Rat
J
2314
EcoRI
ori
EcoRI
EcoRI
major
alternative
NcoI
J+340
EcoRI
~J+1400
*
* —116
* —— 230

cells might have resulted from mutation of the Py sequences in the region of the site, and consequent increased cleavage efficiency. This possibility was investigated by determination of the DNA sequence surrounding the 53-Rat alternative site. This region is contained within the λgt-WES clone 53L (Ruley *et al.*, 1987) which corresponds to the central 53-Rat *Eco*RI fragment (Fig. 3.2); the sequence was determined from the right-hand *Eco*RI site to a position 85 nts upstream from the alternative poly(A) site. No differences from the Py A2 wild-type sequence (Griffin *et al.*, 1981) were detected. The increased frequency of alternative site polyadenylation is therefore more likely due to the nature of the cellular sequences found downstream from the 53-Rat Py insert.

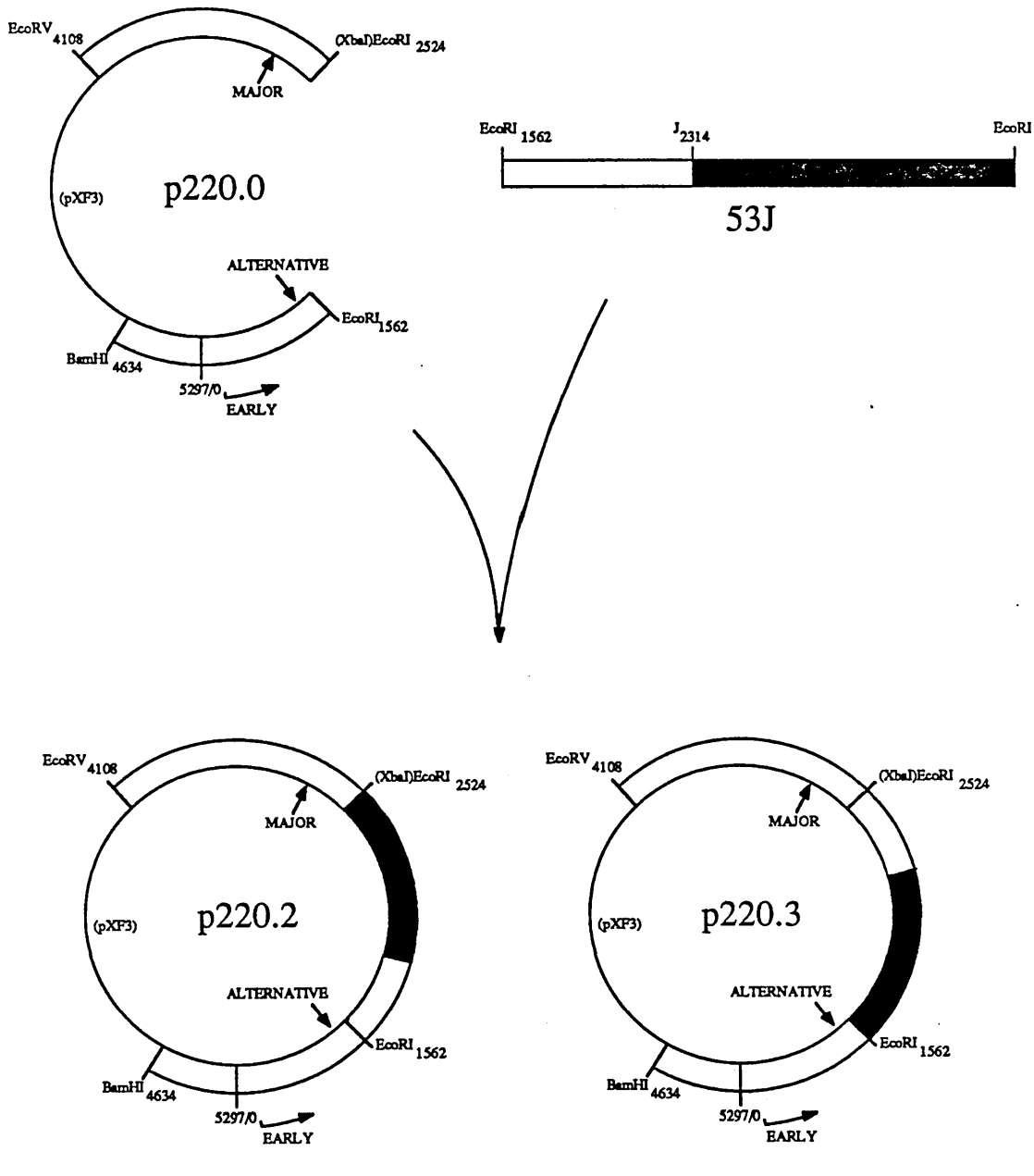
Reconstruction of the 53-Rat polyadenylation pattern in a transient expression system.

The preferential use of the viral alternative poly(A) site in 53-Rat cells might result from blockage of transcription through to an efficient cellular poly(A) site by sequences located between such a site and the viral insert. If this is so then these sequences could be present within the 53J clone. This idea was initially investigated by ligating the 53J *Eco*RI fragment into the *Eco*RI site of vector p220.0 (Fig. 3.5). The vector was designed such that, after insertion of the 53J fragment in the transcriptional orientation in which it was isolated (giving p220.2), the active transcription unit of 53-Rat was reconstructed. Significantly, while the nature of sequences in 53-Rat downstream from the right hand 53J *Eco*RI site is unknown, there is a known, efficient poly(A) site, the Py early major site, in this position in p220.2 (Fig. 3.5).

If the rat sequences contained within p220.2 have no effect on transcription or RNA processing, it would be predicted that transcripts initiated at the Py early cap sites would extend through to, and be predominantly cleaved at, the major Py poly(A) site. This would be exactly analogous to the situation observed in wild type Py lytic infection, when alternative site polyadenylation accounts for at most 15% of the total (see chapter 2). Any

Figure 3.5.

Plasmid reconstruction of the 53-Rat active transcription unit. Plasmid p220.0 was derived from a *Bam*HI-*Eco*RV clone of Py sequences in the plasmid vector pxf3. Py sequences are denoted by white boxes, and plasmid sequences by the thin line. The vector consequently contains all of the Py early region expression sequences. Conversion of the *Xba*I site at Py nt. 2524 to an *Eco*RI site by linker addition was followed by removal of the viral sequences between nts 1562 and 2524 by digestion with EcoRI. The positions of the early Py promoter and the major and alternative poly(A) sites are indicated. The 2.2 kb *Eco*RI fragment 53J is derived from the right hand side of the 53-Rat transcription unit as shown in Fig. 3.2, and spans the viral-host junction at Py nt. 2314. The black box represents 53-Rat cellular DNA adjacent to the site of viral integration. Insertion of this fragment into *Eco*RI-cleaved p220.0 yielded plasmids p220.2 and p220.3. In p220.2 the orientation of the 53J sequences is such that the organisation exactly reproduces that of 53-Rat between the early promoter and the promoter-distal *Eco*RI site. The plasmid differs significantly from 53-Rat, however, in that a known efficient poly(A) site is positioned downstream from the cellular *Eco*RI site.

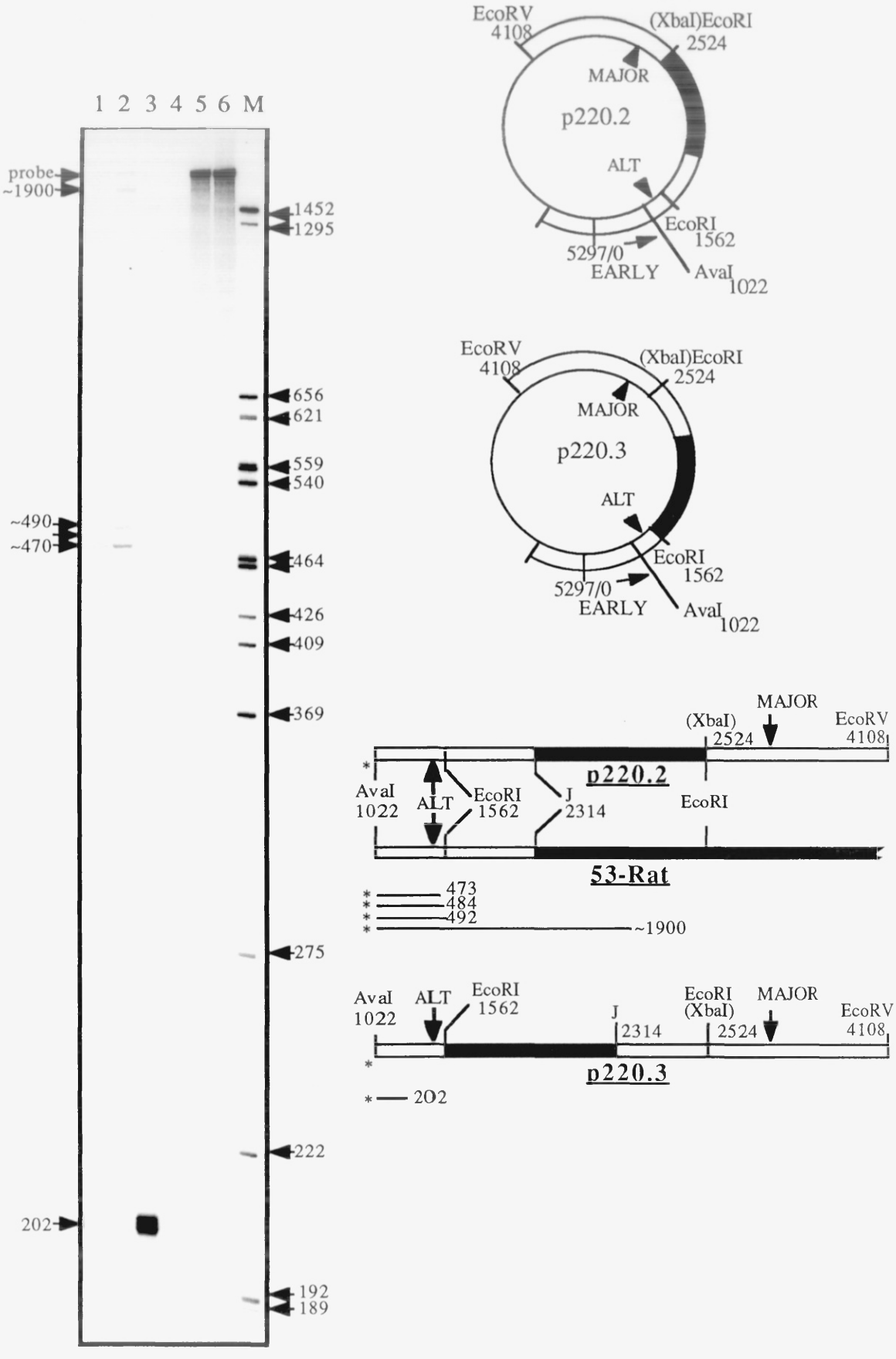

EcoRV 4108
(XbaI)EcoRI 2524
MAJOR
(pXF3)
p220.0
ALTERNATIVE
EcoRI 1562
BamHI 4634
5297/0
EARLY
EcoRI 1562
J 2314
EcoRI
53J
EcoRV 4108
(XbaI)EcoRI 2524
MAJOR
(pXF3)
p220.2
ALTERNATIVE
EcoRI 1562
BamHI 4634
5297/0
EARLY
EcoRV 4108
(XbaI)EcoRI 2524
MAJOR
(pXF3)
p220.3
ALTERNATIVE
EcoRI 1562
BamHI 4634
5297/0
EARLY

one of a number of events directed by the 53J cellular sequences might cause the pattern of polyadenylation of p220.2 transcripts to change. For example, if transcription were to terminate or if primary transcripts were to be cleaved within the cellular sequences, then such transcripts could only become polyadenylated at the Py alternative site, since this would be the only available site on the capped nascent transcripts. In a less extreme case, transcription might pause within, or proceed rather slowly through, the 53J cellular sequences. In this case, the polyadenylation machinery would have more time in which to recognise and cleave the alternative poly(A) site before the transcription complex reached the distal, major site. Any of the events described (hereafter referred to collectively as "transcriptional interference" events) might be predicted to result in preferential alternative poly(A) site use - the 53-Rat phenotype. As a preliminary control to check the orientation dependence of any effect which might be observed, the 53J fragment was inserted into p220.0 in the reversed orientation, giving p220.3 (Fig. 3.5).

Both p220.0 derivatives were transfected into Rat-1 cells, and the polyadenylation patterns of RNAs isolated 48 hours later were compared with that of the parental cell line, 53-Rat (Fig. 3.6). For each of the two plasmids a 3' end-labelled probe was constructed which encompassed both Py poly(A) sites and the rat sequences located between them; the position of labelling was in each case an *Ava*I site at Py nt. 1022. The probe derived from p220.2 was also annealed with total cell RNA from 53-Rat; this probe was contiguous with the 53-Rat sequences as far as the promoter-distal *Eco*RI site (Fig. 3.6). 53-Rat total RNA predominantly protected fragments in the range ~470 to ~490 nts, indicating the preferential use of the alternative poly(A) site. The significant result of this assay was that an essentially identical set of fragments was protected from S1 digestion by RNAs transiently expressed from p220.2; these transcripts thus appeared to be cleaved mainly at the alternative poly(A) site, in spite of the fact that a known efficient poly(A) site was present downstream, just as it would be in the wild-type virus, where the majority of transcripts would use the downstream, efficient site.

Figure 3.6.

Transient expression of p220.2 and p220.3 in Rat-1 cells. Plasmids p220.2 and p220.3 (Fig. 3.5; shown here upper right) were transfected into Rat-1 cells and total cell RNA (40μg) isolated 48 hours after transfection was used in a 3' S1 nuclease assay. Two 3' end-labelled probes were constructed, one from p220.2 and the other from p220.3. Each was labelled at the Py *Ava*I site (nt. 1022) and extended through both poly(A) sites to the Py EcoRV site (nt. 4108). The probes are represented diagrammatically (lower right) with the p220.2 probe aligned with the viral and adjacent cellular sequences of 53-Rat in order to demonstrate homology. Lane 1 contains S1-resistant products obtained after annealing RNA transiently expressed from p220.2 with the p220.2 probe. Lane 2: 53-Rat RNA (p220.2 probe). Lane 3: RNA transiently expressed from p220.3 (p220.3 probe). Lane 4: yeast RNA negative control (probes p220.2 and p220.3). Lane 5: p220.2 probe (no S1). Lane 6: probe p220.3 (no S1). The essentially identical protected fragments in lanes 1 and 2 correspond to RNA 3' ends in the region of the alternative poly(A) site (~470 to ~490 nts) and a low level of apparent RNA 3' ends approximately 600 nts downstream from the viral/host junction . The 202 nt. (±2 nts) band in lane 3 indicates apparent 3' ends at Py nt. 1224 (±2 nts). The protected fragments are represented below the probe diagrams. For the alternative poly(A) site, sizes of protected fragments predicted from the results described in chapter 2 are shown.


1 2 3 4 5 6 M
probe
~1900
1452
1295
656
621
559
540
~490
~470
464
426
409
369
275
222
202
192
189
EcoRV
4108
(XbaI)EcoRI
2524
MAJOR
p220.2
ALT
EcoRI
1562
5297/0
EARLY
AvaI
1022
EcoRV
4108
(XbaI)EcoRI
2524
MAJOR
p220.3
ALT
EcoRI
1562
5297/0
EARLY
AvaI
1022
MAJOR
(XbaI)
2524
EcoRV
4108
p220.2
AvaI
1022
ALT
EcoRI
1562
J
2314
EcoRI
53-Rat
473
484
492
~1900
AvaI
1022
ALT
EcoRI
1562
J
2314
EcoRI
(XbaI)
2524
MAJOR
EcoRV
4108
p220.3
202

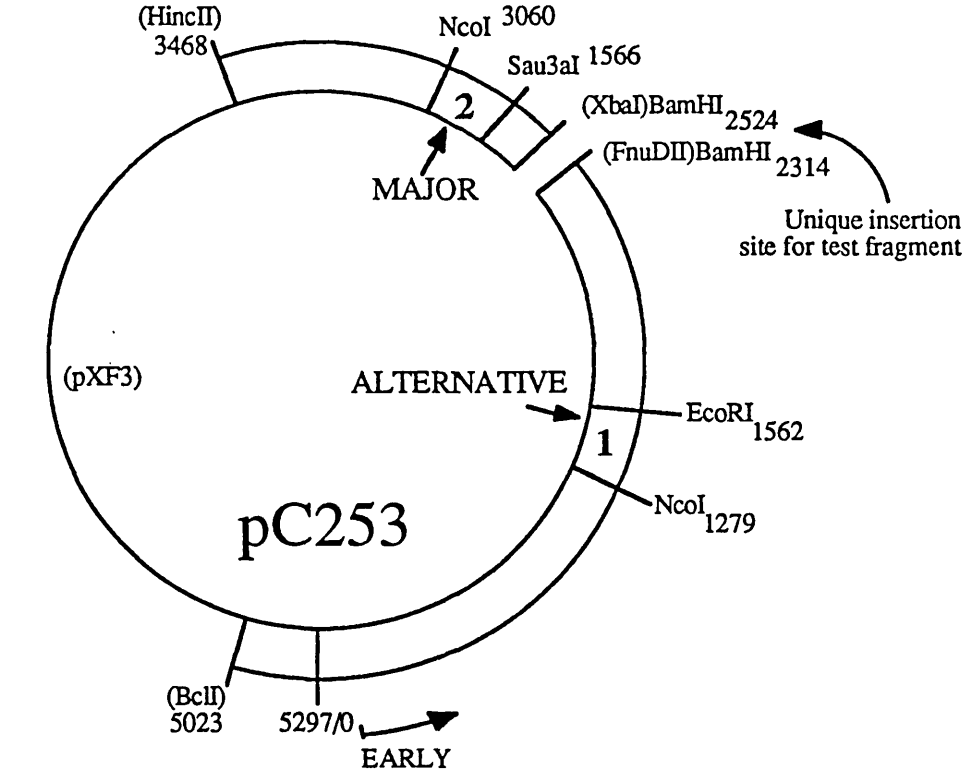
Unexpectedly, transiently expressed RNAs from p220.3 protected a single fragment of ~202 nts (Fig. 3.6), corresponding to an apparent 3' end at Py nt. 1224 (±2nts); this does not correspond to a known Py processing site. Further investigation of this apparent 3' end demonstrated that it is, rather, a novel 5' splice site which appears to become activated upon insertion of the inverted 53J cellular sequences at the downstream position in p220.3. This novel splice is discussed further in a later section of this chapter.

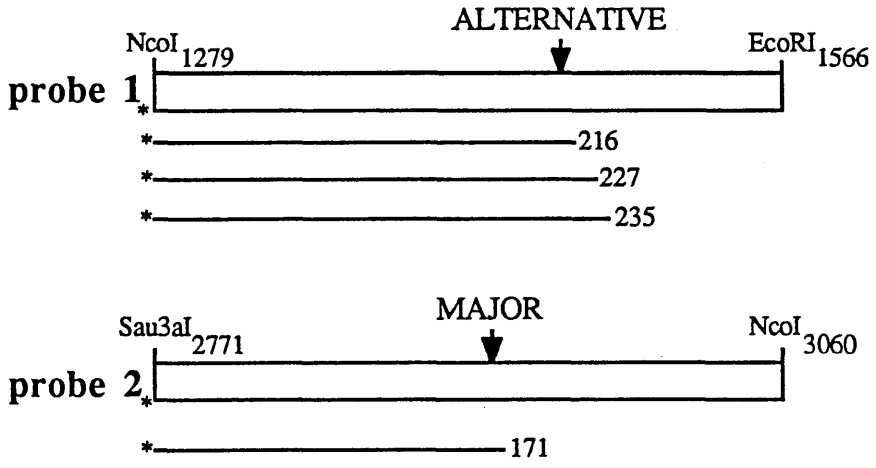
Having shown that preferential use of the alternative poly(A) site could be conferred on a Py plasmid by insertion of cloned sequences from 53-Rat, a second assay system was devised (Fig. 3.7), with a view to defining more accurately the sequence element(s) responsible for this effect. The vector, pC253, is essentially a version of p220.2 with all of the rat sequences removed; these sequences were replaced by a unique *Bam*HI cloning site. pC253 retains, as a consequence, all of the Py sequences between the transcriptional control region and the position of the 53-Rat viral-host junction at nt. 2314 (fortuitously the position of a *Fnu*dII restriction site). As with p220.2, the Py major early poly(A) site is present just downstream from the site at which fragments of rat DNA were inserted.

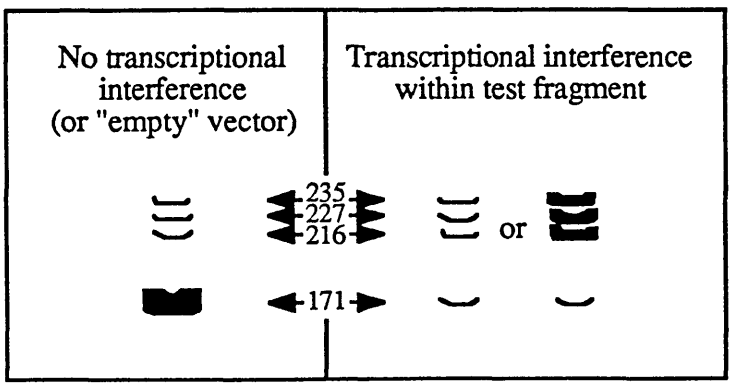
After insertion of test fragments into pC253 and transient expression of the resulting plasmids in Rat-1 cells, two 3' end-labelled probes were annealed simultaneously with total cell RNA prior to S1 digestion. The two probes span the alternative and major poly(A) sites respectively (Fig. 3.7); since they contain only pC253 sequences, the probes could be used in annealing with RNAs transiently expressed from any pC253 derivative. The alternative site probe was labelled at the *Nco*I site at Py nt. 1279; transcripts polyadenylated at the alternative site were therefore predicted to protect a triplet of fragments of 216, 227 and 235 nts (see chapter 2). The major site probe, being labelled at the *Sau*3aI site at Py nt. 2771, was predicted to be protected by transcripts cleaved at the major site to give a single band of 171 nt. Comparison of the intensity of the 216-235 nt. triplet to that of the 171 nt. band within each track of the autoradiograph would therefore give an indication of the polyadenylation pattern for a given plasmid. Furthermore, this

Figure 3.7.

A plasmid-based assay for sequences which prevent read-through RNA formation. The plasmid vector pC253 was designed such that a unique cloning site (BamHI) is present at precisely the point where viral sequences have recombined with cellular DNA in 53-Rat (see Fig. 3.2). The *Bcl*I-*Hin*cII fragment of Py A2, including the entire viral early region and spanning nts 5023-3468, was cloned into pXf3. Sequences between the *Fnu*DII site at nt. 2314 and the *Xba*I site at nt. 2524 were then replaced with a *Bam*HI linker. After insertion of a test fragment at the *Bam*HI site, pC253 derivatives were transiently expressed in Rat-1 cells. 40 μg total cell RNA was then used in a 3' S1 analysis simultaneously employing two probes (1 and 2, represented in the centre of the diagram). Probe 1 spans the alternative Py poly(A) site and is protected by mRNAs cleaved at this site to give fragments of 216, 227 and 235 nts, as predicted by the results described in chapter 2. Probe 2 was designed to give a protected fragment of 171 nts when annealed with mRNAs cleaved at the Py major poly(A) site. At the bottom of the figure is a representation of predicted autoradiograph patterns which would be obtained using probes of equal specific activities. The first example is the result expected if pC253 or a derivative carrying an inert test fragment is used: the majority of the signal corresponds to major site polyadenylation, just as when wild-type Py early mRNA is used (Fig. 2.4). If transcription terminates, or if primary transcripts are cleaved, within the test fragment, then either of the patterns shown on the right might be expected (for a more detailed discussion see text).


(HincII)
3468
NcoI 3060
Sau3aI 1566
2
(XbaI)BamHI 2524
(FnuDII)BamHI 2314
MAJOR
Unique insertion
site for test fragment
(pXF3)
ALTERNATIVE
EcoRI 1562
1
NcoI 1279
pC253
(BclI)
5023
5297/0
EARLY


ALTERNATIVE
NcoI 1279
EcoRI 1566
probe 1
216
227
235
MAJOR
Sau3aI 2771
NcoI 3060
probe 2
171


No transcriptional
interference
(or "empty" vector)
Transcriptional interference
within test fragment
235
227
216
or
171

ratio could be standardised by comparison with the signals observed for RNAs transiently expressed from the vector, pC253. This assay could thus be used to detect any shift in poly(A) site preference conferred by fragments of rat DNA inserted into the *Bam*HI site of pC253 (see Fig. 3.7 for theoretical examples).

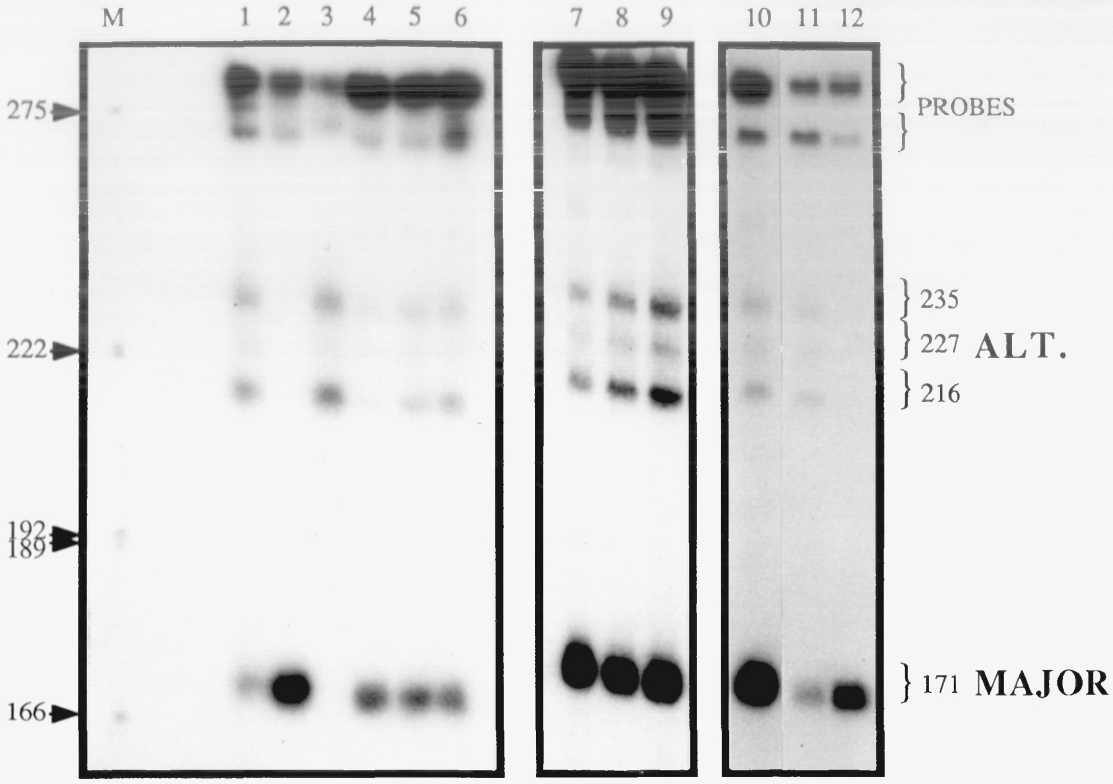
Four fragments of the cloned 53-Rat sequences were tested in both possible orientations in the pC253 assay (for plasmid designations see Fig. 3.8). A 'transcriptional interference' effect had appeared to cause a shift to favour the alternative poly(A) site of p220.2 (Fig. 3.6). If this effect could be directed by a simply-defined section of the 53J cellular sequences up to a few hundred nucleotides in length, then the sub-fragments chosen for pC253 analysis (Fig. 3.8) were such that one of them should contain this element. Since the specific activity of each probe was not exactly the same in different experiments, a pC253 internal standard was included in each assay.

After densitometric scanning of the autoradiographs (Fig. 3.9), the peak integrals were used to calculate a value for the intensity of the alternative poly(A) signal, relative to that of the major poly(A) signal, for each pC253 derivative. Inclusion in the calculations of the values obtained using the 'empty' vector made possible the comparison of results obtained in different experiments. The value 'X' for the standardised, relative use of the alternative poly(A) site is described by a simple formula given in the legend of Fig. 3.9. If the test fragment inserted into pC253 has no effect on the polyadenylation pattern of transiently expressed RNAs, then the ratio of alternative: major site signal will be the same as that seen with pC253; the value of 'X' in this case would be 1.0. If the fraction of RNAs cleaved at the alternative site is increased by the insertion of a test sequence, then the value of 'X' will be greater than 1. Conversely, values of 'X' less than 1 would be expected if polyadenylation at the alternative site is disfavoured.

The results presented in Figs 3.8 and 3.9 serve to confirm and extend the conclusions drawn from the transient expression study using p220.2 (Fig. 3.6). Thus pCFR1, a pC253 derivative containing all of the cellular sequences of p220.2 in their original

Figure 3.8.

pC253-based deletion analysis of the 'transcriptional interference' sequences from 53-Rat. The fragments of rat DNA tested in the pC253 assay (Fig. 3.7) are represented in the lower part of the diagram. Each fragment was inserted into pC253 in both transcriptional orientations, resulting in the plasmids designated as shown. pCFR1 and pCFR2 contain the rat sequences from the *Fnu*DII site at the viral/host junction (J) to the right-hand *Eco*RI site. pCAV1 and pCAV2 contain the internal *Ava*II fragment indicated. ΔA1 denotes the position of a Bal-31 deletion end-point, which marks the rightward extent of sequences contained in pCΔA1 and pCΔA2. This end point is located 861 nts downstream from the viral-host junction. pCBgR1 and pCBgR2 contain sequences between the *Bgl*II site and the right-hand *Eco*RI site. The autoradiographs obtained after performing three separate experiments (using the strategy described in Fig. 3.7) are shown in the upper half of the figure. The lane number corresponding to each test fragment is indicated to the right of the plasmid designation. Lanes 1 and 2: pCFR1 and pCFR2. Lanes 5 and 6: pCAV1 and pCAV2. Lanes 8 and 9: pCΔA1 and pCΔA2. Lanes 11 and 12: pCBgR1 and pCBgR2. Lane 3 shows the signal for 53-Rat total RNA (40μg) and lanes 4, 7 and 10 correspond to the pC253 (empty vector) signal for each experiment.


M
1 2 3 4 5 6
7 8 9
10 11 12
275
222
192
189
166
PROBES
235
227 ALT.
216
171 MAJOR

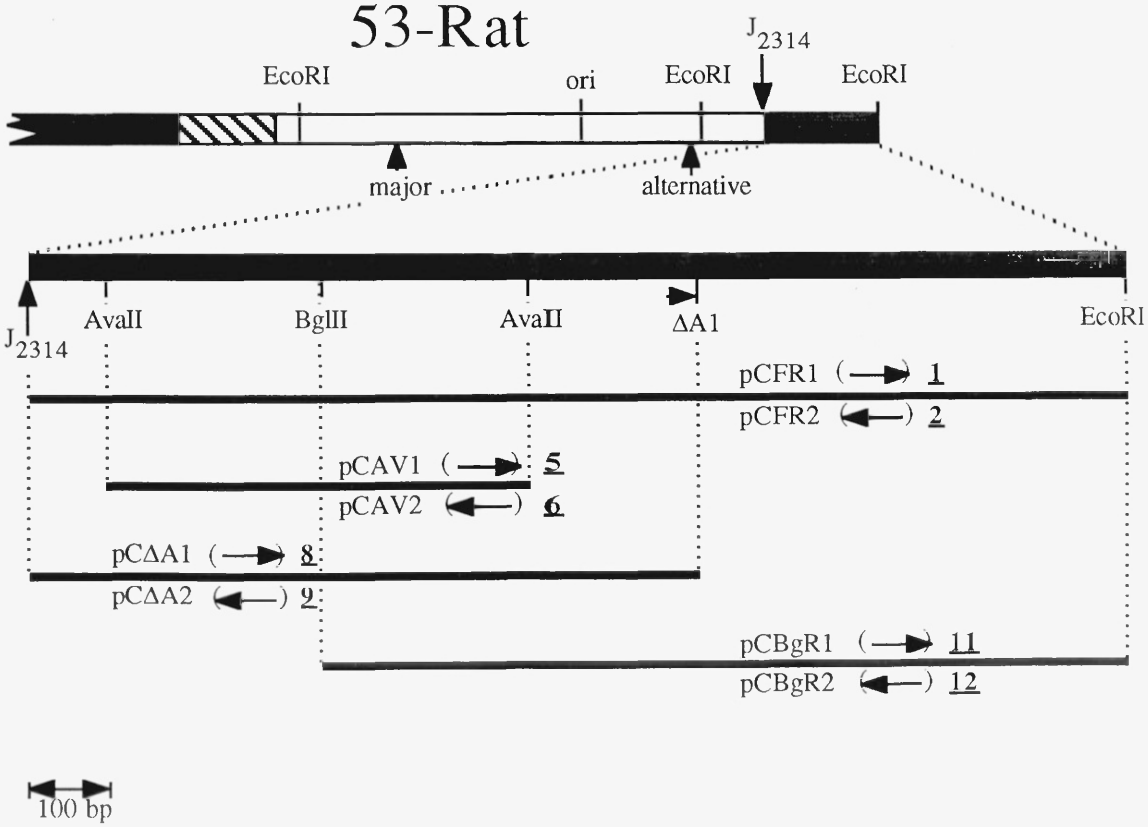

53-Rat
J2314
EcoRI
ori
EcoRI
EcoRI
major
alternative
J2314
AvaII
BglII
AvaII
ΔA1
EcoRI
pCFR1 (→) 1
pCFR2 (←) 2
pCAV1 (→) 5
pCAV2 (←) 6
pCΔA1 (→) 8
pCΔA2 (←) 9
pCBgR1 (→) 11
pCBgR2 (←) 12
100 bp

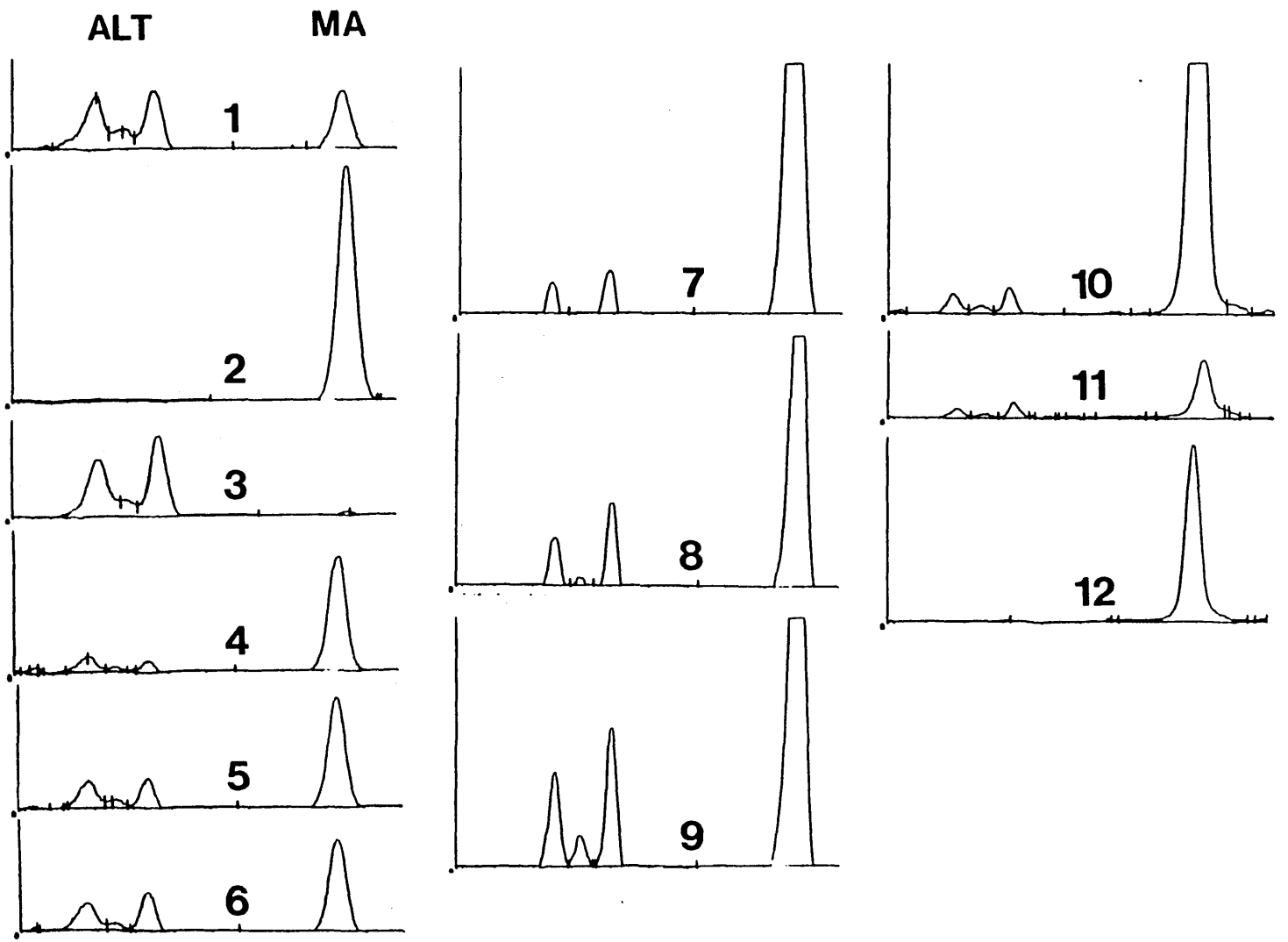
Figure 3.9.

Densitometric analysis of the data presented in Fig. 3.8. A Joyce-Loebl Chromoscan 3 densitometer was used to scan lanes 1-12 of Fig. 3.8. Peaks corresponding to the alternative (ALT) and major (MA) poly(A) site signals are indicated for the scan of lane 1. Results from the three separate experiments can be compared after normalising for the pC253 (empty vector) signal. A unitless value X for the relative usage of the alternative poly(A) site is given by:

$$X=(ALT_T x MA_{EV})/(ALT_{EV} x\ MA_T)$$

where ALT_T and ALT_{EV} are the integral values of alternative poly(A) signal for the test plasmid and the empty vector respectively, and MA_T and MA_{EV} are the corresponding values for major poly(A) signal.

LANE	PLASMID	ALT_T	MA_T	ALT_{EV}	MA_{EV}	X
1	pCFR1	11906	4888	2436	10156	10.16
2	pCFR2	38	20202	"	"	0.01
3	(53-Rat)					
4	pC253	(2436)	(10156)	"	"	1.00
5	pCAV1	4840	9062	"	"	2.23
6	pCAV2	6128	7836	"	"	3.26
7	pC253	(3298)	(27916)	3298	27916	1.00
8	pCΔA1	6162	22108	"	"	2.36
9	pCΔA2	12902	25604	"	"	4.27
10	pC253	(3274)	(30350)	3274	30350	1.00
11	pCBgR1	1658	5432	"	"	2.83
12	pCBgR2	44	14146	"	"	0.03


ALT
MA
1
2
3
4
5
6
7
8
9
10
11
12

transcriptional orientation, is transcribed to give an X value of 10.16 (Fig. 3.8). This gives a measure of the maximal extent to which alternative site polyadenylation is favoured in this system. Perhaps surprisingly, none of the three internal sub-fragments of the 53J sequence retained full activity when tested in the pC253 assay. Thus plasmids pCAV1 (containing an internal *Ava*II fragment), pCΔA1 (containing the same *Ava*II fragment with additional rat sequences at each end) and pCBgR1 (containing a *Bgl*II-*Eco*RI fragment substantially overlapping the fragments used for pCAV1 and pCΔA1) all gave intermediate levels of alternative site polyadenylation when compared with pC253 (the empty vector) and pCFR1 (containing the entire 1.4 kb of rat DNA from the 53J clone; Fig. 3.9). In fact, when the inserts of pCAV1 and pCΔA1 (X values 2.23 and 2.36) were reversed, to give pCAV2 and pCΔA2 (Fig. 3.8), the relative levels of alternative site polyadenylation were increased (X values 3.26 and 4.27, respectively; Fig. 3.9). In contrast, reversal of the fragment orientation of pCFR1 and pCBgR1, giving pCFR2 and pCBgR2 (Fig. 3.8) caused a marked decrease in the X values (0.01 from 10.16 and 0.03 from 2.83, respectively; Fig. 3.9). This decrease in alternative poly(A) site use is the result of activation, in primary transcripts derived from these plasmids, of the novel splice donor identified in Fig. 3.6 and investigated in further detail in a later section of this chapter. The sequences removed by this splicing event span the Py alternative poly(A) site, which is consequently under-represented in the steady-state RNA.

Cellular DNA sequences directing Py alternative site polyadenylation.

The polyadenylation phenotype seen in 53-Rat (Figs 2.5, 3.2) and in the plasmid reconstructions p220.2 and pCFR1 (Figs 3.6, 3.8) could be due to POLII termination or pausing within the cellular DNA of the 53J clone. In order to make comparisons with known POLII terminators, the DNA sequence was determined from the 53-Rat viral-host junction to the *Eco*RI site at the 3' extent of the cloned fragment (Fig. 3.10). This

Figure 3.10.

The DNA sequence downstream from the 53-Rat viral insert. The sequence shown extends from the right-hand viral-host junction to the distal *Eco*RI site (see Fig. 3.2). Cleavage sites for the restriction enzymes shown to the right are underlined in the sequence. Also underlined are a poly(pyrimidine) stretch extending from 456 to 494 nts downstream from the junction, and two copies of the AATAAA hexanucleotide beginning at nts 512 and 551. The significance of these features is discussed in the text.

J+1

```
⬇      10         20         30         40         50         60         70         80         90     AvaII
CGGGGACAGC AGTGCCAGGA AACAAAAGCA GGGAGCTGAT TTGATTGAGC TCCCTGACAG GTTTGAAGGA ATAGACTAGC TTGCCAAGAA CATTCGAGGG
      110        120        130        140        150        160        170        180        190        200
TCCTGTGCCA CCATCCAGGC TCCAATTAAA CCCCATGAAG CCTATTCTCC TGTTCTTAGG TAACAATTCT CAATCATAGA ATCTATTGCT ACATGAAAAC
      210        220        230        240        250        260        270        280        290        300
CCAAAGGAAC ATGGACAGTA TTTACAGGGA AAAGAAATTA TAACACACAC CAGGAGATTG CCATGTGTGT CCACCTGAGC AGATGTGAAA TGTTGTACTG
      310        320        330   NcoI             350        360        370        380     BglII         400
CTCAAAGTGA GGCTCTCAGC CACATCATTA TCATGCCATG GAAATTGACA TAGGATAATT CTCAGGCCAA ATCAGAACAC AGAATACCAG ATCTTAAAGT
      410        420        430        440        450        460        470        480        490        500
TATACACACA AACGCACTCA CACATGCATA CACACACACA AACTCACACA AAACACTCTC TCTCTCTCCC TCCCTCTCTC TCTCTCTCTC TCTCATACAA
      510        520        530        540        550        560        570        580        590        600
TTAGAATGCC AAATAAATAC GGAGGTGTGG CATTATCTCA AAATACAATC AATAAAATCA GAGAAAAGCA ATGATACCCT GTTGACAGTG CAGTCACACA
      610        620        630           AvaII  650        660        670        680        690        700
CTGTGAGGCA GCAAGAGGGA TTAGGCCCCT CCAGCTCTTC ACAGGACCTA TTGAACCCAA AGCACCTCCA TGTCAATTGC AACCAATGAT TGCTCATAAC
      710        720        730        740        750        760        770        780        790        800
CAGAGCACCT AGAAATCTCC CACAATCCAA CAACTGTCCA GGCTACTTAA AGCAACACAA TGAGAGCACA CACTCTACTA CCCTTATTCC CAGAGCTGTA
      810        820        830        840        850        860        870        880        890        900
ATTCAGGCCA TAGGCTATGA TTTAGCGTTT GTAGGAATCA AAATAGCCTG TGCCACCCTC TCATTCCCAT CTGATCTTCA CATTCTGATG CAATCAGGAA
      910        920        930        940        950        960        970        980        990       1000
AGTGATTTTG ACCATGAGCC CATCATGGAG TTGTAGCCTA CTGTGATGTG AGGAAATGTG GATTTAAACA GAACTTGCAC TGTGACTACC TGTCATTTAA
     1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATTCTTTCCT GGGTAATCCG CCAGGATTCT CCTGAGTTCA TCTCTCTCCT TTTGCATCCT TGTGGAGGGC AATTTTGAGG TTCTCCAAAC GCTTCATTCT
     1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
CTCCTCCTCA ACAGCTGTCG TGGAGGGTTG GGATGATGCC TTCTGATCAG ACTCTGTAGG TAGATAAACA CCAGTAAACA GGCATATATG GGGACATAAT
     1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GTCAATGAAA AACTATGCTT ACTCTAAAAC AGAAGTCTGA GGATGAACAC TTCTACAGAT ACAGTGACTC CCTGACAGAA CAAGAAAGCT ATCTTTAAGC
     1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
TGAACTCAGC TCAGTCCCTG TTCCCCACCA TCATAGACAT ATGGTGACAT ATGCCCATCA TAGATATAGA CCACACCTCC TACAGCCATC CTTCGTGCAC
     1410       1420      EcoRI
TGAAATTCTA AAGCTGGCCT AAAGAAGAAG AATTC
```

SITE	ENZYME	
99	Ava II	[G\G{AT}CC]
336	Nco I	[C\CATGG]
389	Bgl II	[A\GATCT]
644	Ava II	[G\G{AT}CC]
1430	EcoR I	[G\AATTC]

sequence contains a number of interesting features which relate to the results obtained by 3' S1 mapping. A poly(pyrimidine) stretch $(CT)_6C_3TC_3(TC)_{10}$ extends between 456 and 494 nts downstream from the viral-host junction. The promoter-proximal end of the poly(pyrimidine) stretch corresponds exactly to the position of apparent 3' ends in 53-Rat poly(A)$^+$ and poly(A)$^-$ RNAs defined by the 116 nt. protected fragment shown in Fig. 3.4. The most likely explanation for the generation of this fragment is that the (CT) repeat allows slippage of the RNA-DNA duplex, transiently generating single-stranded loops which are subject to S1 cleavage. The ability of a (CT) repeat to cause S1 hypersensitivity in RNA-DNA duplexes has been reported previously (Johnson *et al.*, 1986). Even though there are probably no genuine RNA 3' ends in the vicinity of the (CT) repeat, the generation of an S1 signal at this position indicated that a small proportion of the steady-state 53-Rat transcripts extend beyond the repeat. A second protected fragment some 230 nts in length (Fig. 3.4) indicated the presence of a low level polyadenylated 3' end in the region of 570 nts downstream from the viral-host junction. Examination of the DNA sequence in this region showed that this position is 14 nts downstream from the second of two AATAAA hexanucleotides found in the 53J sequence (Fig. 3.10).

Computer analysis of the 53-Rat downstream sequence revealed no dyad symmetry elements significantly more extensive than those found within the Py early region. Dyad symmetry elements might, in theory, generate hairpin loop structures in a template DNA and/or its RNA transcript, and such structures have been implicated in transcriptional pausing and premature termination in both prokaryotes (Yanofsky, 1981) and eukaryotes (Ben-Asher and Aloni, 1981; Hay and Aloni, 1984; Maderious and Chen-Kiang, 1984).

Comparison of the 53-Rat sequence with the sequences of elements implicated in termination of POLII transcription downstream from the sea urchin histone H2A and mouse β_{major}-globin genes (Johnson *et al.*, 1986; Citron *et al.*, 1984) revealed four stretches of homology. The transcript-sense strand of the H2A terminator "3' box" shares a 17/25 nt. homology with the 53-Rat sequence:

```
H2A 3' BOX    96    TATATGGttatCATAAT  CAtTGcAA    120
53-Rat      1185    TATATGG ggaCATAATgtCAaTG AA   1209
```

When the opposite strand of the H2A 3' box sequence was used, two further stretches of homology (22/28 and 15/19 nts) with the 53-Rat sequence were detected:

```
H2A 3' BOX  12    ATgAtAAC CA tATACCAG TCTTAAA     36
53-Rat     371    ATcAgAACaCAgaATACCAGaTCTTAAA    398


H2A 3' BOX  81    TGtGCA AaGTtAAATGTT      98
53-Rat     276    TGaGCAgAtGTgAAATGTT     294
```

For the β_{major}-globin terminator, one significant (18/23 nts) homology with the 53-Rat sequence was found:

```
Globin      8    AGAAtAAGAAtAtCTAgTtTTTAAG     32
53-Rat   1277    AGAAcAAGAA AgCTA TcTTTAAG   1299
```

This section of the β-globin element is located near the 5' end of the fragment 'F' described by Citron *et al.* (1984).In addition, the computer search identified homologies largely based on the (CT) repeat sequence; a similar repeat is found downstream from the H2A gene and, in a shorter and more degenerate form, downstream from the mouse β_{major}-globin gene. It should be stressed that, while the homologies shown were more extensive than those which would be expected in random sequences of equivalent lengths, this homology search was directed specifically at the two terminator sequences shown. No extensive homologies with the 53-Rat sequence were found in global database searches.

Location of the novel 5' and 3' splice sites used by p220.3 transcripts.

An unexpected complication of the plasmid reconstructions of 53-Rat had been the efficient activation, in certain plasmid constructions, of a novel Py splice donor in the

vicinity of Py nt. 1224 (Figs 3.6, 3.8). The sequence of this site is similar to those of the previously described Py 5' splice sites (see also Table 1.2):

Sequences at Py splice donors:

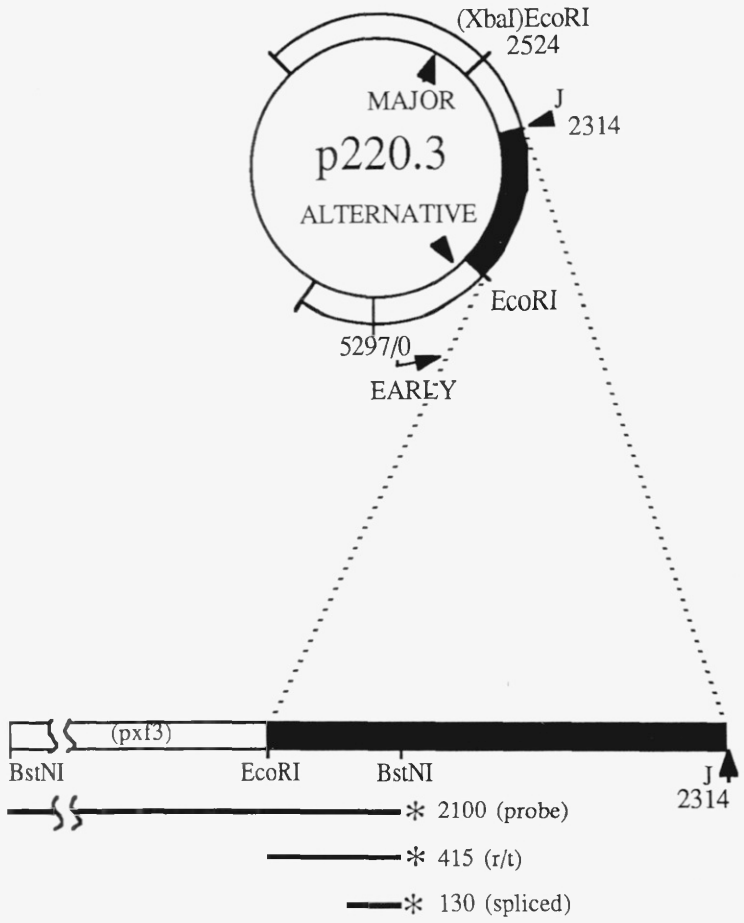
	exon	intron
Novel early, 1224-	ACAA	GTACCCCA
Late, 5022-4124 (VP1)	TCAA	GTAAGTGA
Early, 411-797 (large T-Ag)	CCAG	GTAAGAAG
Early, 748-797 (middle T-Ag)	CCAA	GTAAGTAT

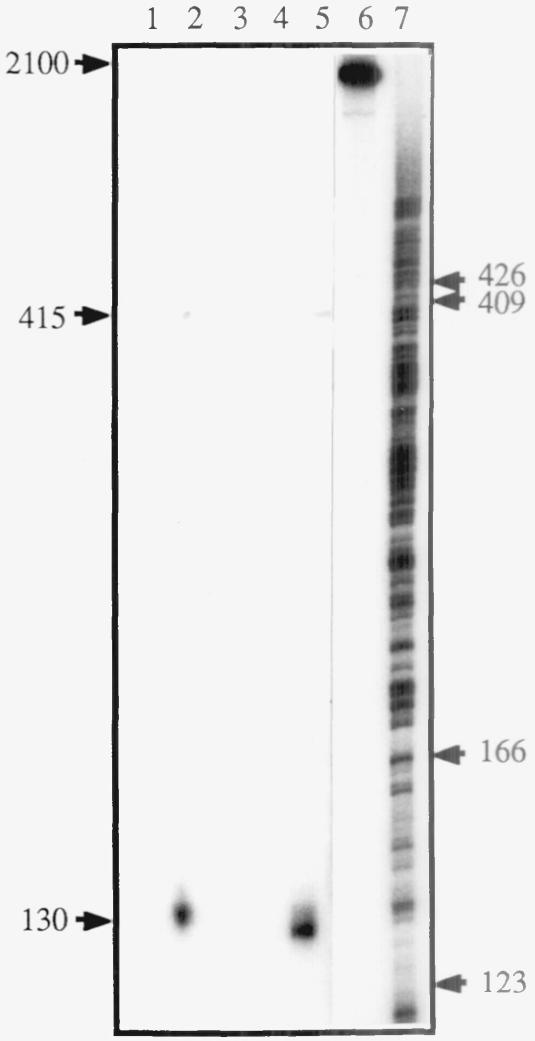
The fact that this 5' splice site was active in plasmids p220.3, pCFR2 and pCBgR2, but not in plasmids pCΔA2 and pCAV2, suggested that cellular sequences located between the Bal-31 deletion endpoint ΔA1 (861 nts downstream from the viral-host junction) and the distal *Eco*RI site (1430 nts downstream from the junction) were responsible for activation of the novel Py 5' splice site. The sequence element involved was considered worthy of further investigation, since the preferential selection of particular pairs of splice sites is poorly understood. A splice acceptor site within the inverted cellular sequences of p220.3 was approximately located in an S1 analysis using a probe 5' end-labelled at the unique *Bgl*II site; this splice acceptor was then mapped more accurately using a probe 5' end-labelled at a *Bst*NI site located 1022 nts downstream from the viral-host junction (Fig. 3.11). This probe was protected by RNAs transiently expressed from p220.3 to give a fragment of 130 nts (±1nt.). The same fragment was protected when RNA from t3.4, a Rat-1 cell line transformed by p220.3, was used. In addition, both RNAs contained a small proportion of unspliced transcripts, which gave rise to a 415 nt. protected fragment. No signal at either position was obtained after annealing the probe with Rat-1 or yeast RNAs. The 130 nt. protected fragment maps an efficient 3' splice site to a position 1152 nts (±1nt.) downstream from the 53-Rat viral-host junction; this site is present only on RNAs whose transcriptional orientation is

Figure 3.11.

Location of the 3' splice site specifically activated in p220.3 transcripts. A single-stranded, 5' end-labelled probe of ~2100 nts was constructed (lower left) by *Bst*NI digestion of a pxf3 clone of the *Eco*RI fragment from 53-Rat spanning the right-hand viral-host junction. Within this probe, sequences between the labelled *Bst*NI site (denoted by an asterisk) at 1024 nts from the viral-host junction and the *Eco*RI site are complementary to transcripts colinear with the p220.3 template (upper left). Such transcripts consequently protect a 415 nt. fragment in S1 analysis. RNAs which are spliced using a splice acceptor between the *Eco*RI and *Bst*NI sites protect the probe to give a fragment of 130±1 nt. The right-hand panel shows the S1-resistant products obtained after annealing the probe with:

Lane 1: RNA from p220.2. Lane 2: RNA from p220.3. Lane 3: yeast RNA. Lane 4: Rat-1 total cell RNA. Lane 5: RNA from t3.4, a cell line derived after transformation of Rat-1 cells by p220.3. Lane 6 contains an aliquot of the undigested probe and lane 7 the products of a (C+T) sequencing track obtained by treating the probe as described by Maxam and Gilbert (1980). The positions of size markers are shown to the right of the panel.


(XbaI)EcoRI
2524
MAJOR
J
2314
p220.3
ALTERNATIVE
EcoRI
5297/0
EARLY
(pxf3)
BstNI
EcoRI
BstNI
J
2314
2100 (probe)
415 (r/t)
130 (spliced)


1 2 3 4 5 6 7
2100
415
130
426
409
166
123

opposite to that of the Py-rat fusion transcripts of the parental line, and does not appear to be represented to any extent in either spliced or unspliced RNAs from untransformed Rat-1 cells (Fig. 3.11). The sequence of this splice acceptor is similar to those of the Py splice acceptors given in Table 1.2:

Sequences at p220.3 and Py splice acceptors:

	intron	exon
p220.3 acceptor	TTTATCTACCTACAG	AGT
Late, 5022-4124 (VP1)	TTCCTTTAATTCTAG	GGC
Late, 5022-5078 (leader)	ATTTTCTATTTTAAG	AGT
Late, 5022-4709 (VP3)	CTACTATTTCCCTAG	GAA
Early, 411-797 (large T-Ag)	CCTATATTCTTACAG	GGC
Early, 748-811 (small T-Ag)	GGGCTCTCCCCCTAG	AAC

Note that while the predominant mRNAs transiently expressed from p220.3, pCFR2 and pCBgR2 do not contain the region of the alternative poly(A) site, they do contain the region between the cellular splice acceptor and the major Py poly(A) site. Furthermore, the latter site is used efficiently (see, for example Fig. 3.8, lanes 2 and 12).

The 53-Rat sequence does not act as a self-contained POLII terminator.

Studies of the possible involvement of RNA processing events, in particular polyadenylation, in the termination of POLII transcription have not given rise to a universally applicable model. Thus, while 3' termination in at least one case required the presence of a functional poly(A) site (Falck-Pedersen *et al.*, 1985), termination at the 3' end of the sea urchin H2A gene did not require prior processing (Birchmeier *et al.*, 1984). A second example of apparent POLII termination in the absence of 3' processing has come from a study of sequences derived from the 3' flank of the human gastrin gene (Sato *et al.*, 1986). These sequences were able to prevent transcription through to a SV40 poly(A) site when inserted into an SV40-CAT vector at a position upstream from the poly(A) site and downstream from the CAT gene. Similarly, a 1008 nt. fragment from the

Ad2 terminal region caused transcription to terminate before reaching the downstream SV40 poly(A) site (Dressler and Fraser, 1987), when the Ad2 fragment was similarly inserted into an SV40-CAT vector.

An approach essentially identical to that used in the latter study was taken using sequences from 53-Rat, in an attempt to establish whether or not these sequences can act to terminate POLII transcription in the absence of an upstream poly(A) site. The vector used for this experiment, pSV2CAT (Gorman *et al.*, 1982), and the derivatives constructed are shown in Fig. 3.12. Two fragments from 53-Rat were inserted in both possible orientations into the unique *Hpa*I site of pSV2CAT, just upstream from the SV40 poly(A) site which is used by CAT transcripts. The larger of these two fragments contained all of the cloned sequences from the 53-Rat 3' flank; the smaller is the *Ava*II fragment which spans the poly(pyrimidine) tract and both AATAAA hexanucleotides (Figs 3.8, 3.10). The four plasmid derivatives and the parental vector, pSV2CAT, were transfected in duplicate into Rat-1 cells, and the resultant CAT activities were determined (Fig. 3.13).

The principal conclusion to be drawn from these data is that in no case is the level of CAT activity greatly depressed below the level obtained with the positive control, pSV2CAT. There was a slight, but reproducible, decrease in CAT activity when the entire cloned 53J fragment was inserted into pSV2CAT in the transcriptional orientation in which it is found in 53-Rat (Fig. 3.13, pCP3), but for the three other derivatives the observed activity is, if anything, enhanced. Thus the 53-Rat 'transcriptional interference' sequences do not appear to act as an efficient terminator of POLII transcription in a situation where there is no upstream poly(A) site.

Figure 3.12.

Construction of pSV2CAT derivatives. The vector pSV2CAT (Gorman *et al.*, 1982) contains the Tn9 chloramphenicol acetyl transferase (CAT) gene driven by the SV40 early region expression signals. A unique *Hpa*I site is situated just upstream from the SV40 early poly(A) site. The 545 bp *Ava*II and ~2.1 kb *Eco*RI fragments derived from 53-Rat and shown in the upper right were inserted, after filling in the recessed 3' termini with Klenow polymerase, in both possible orientations into *Hpa*I-cleaved pSV2CAT. Derivative pCP1 carries the *Ava*II fragment in the same transcriptional orientation as in 53-Rat; this fragment is in the opposite orientation in pCP2. Similarly, pCP3 carries the *Eco*RI fragment in the "correct" orientation, and the fragment is reversed in pCP4.

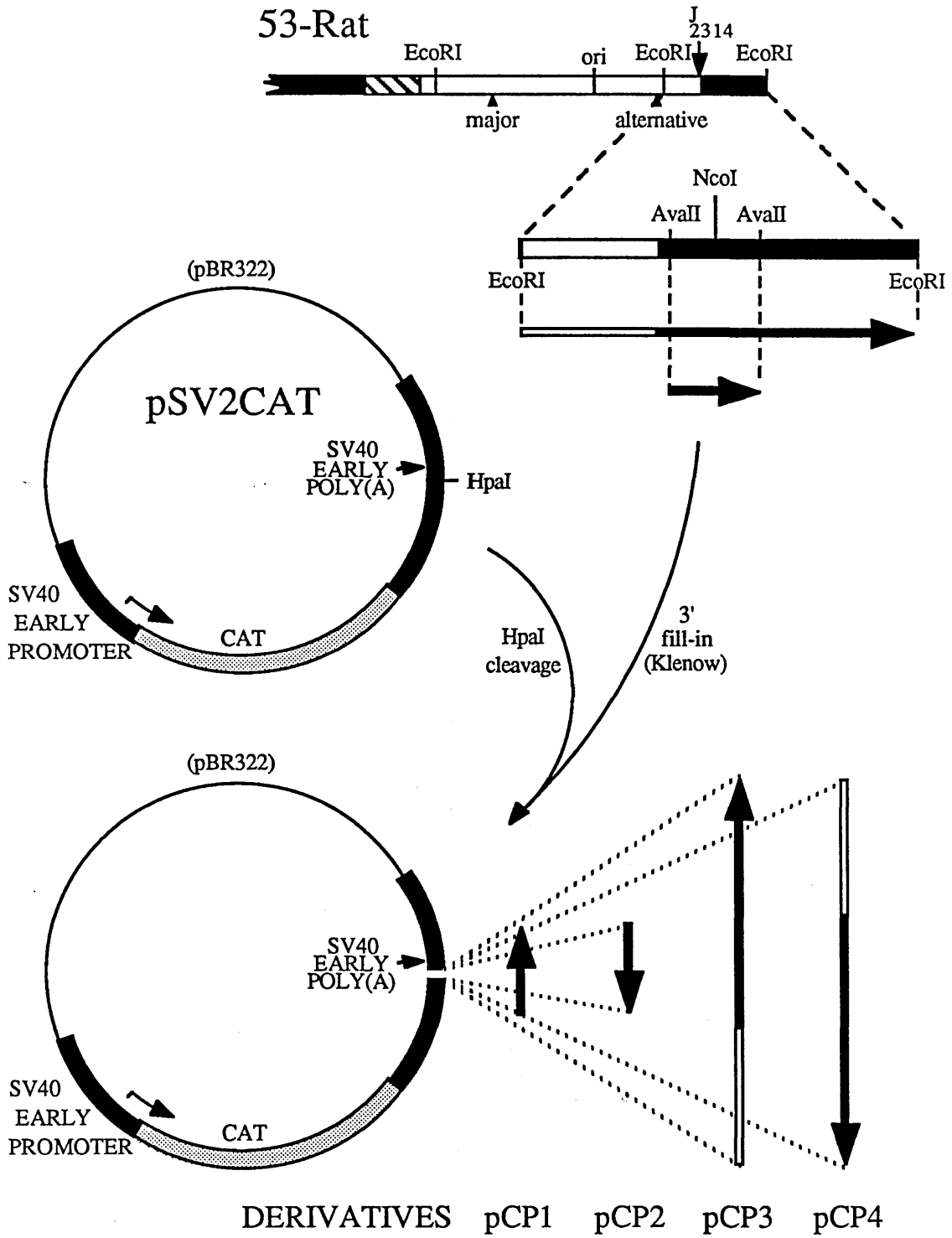
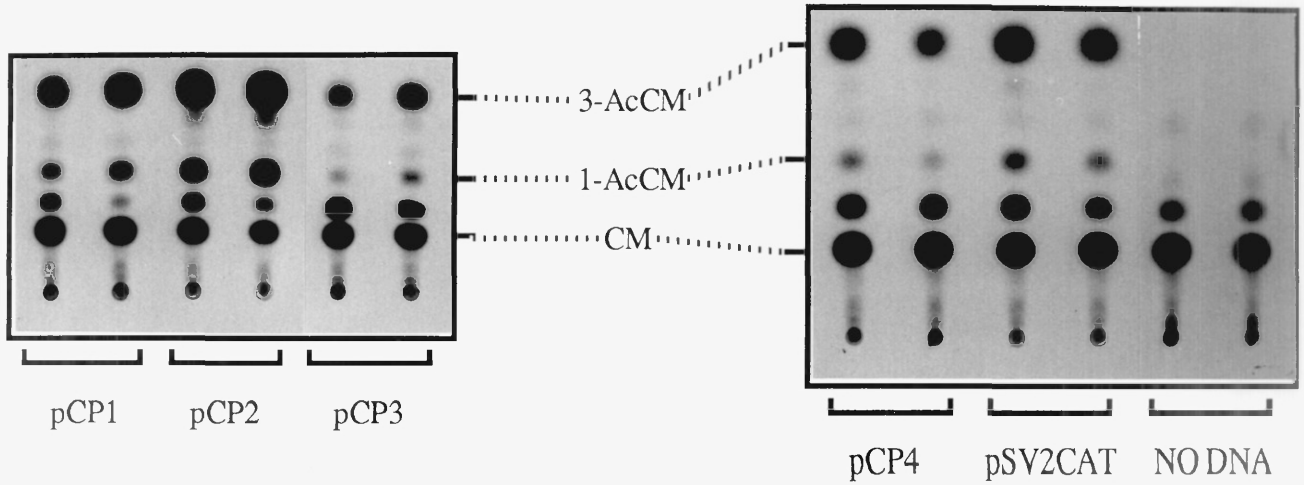
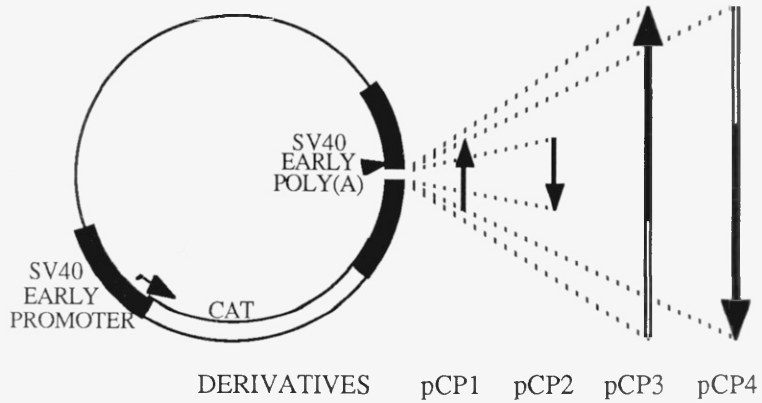

53-Rat
J
2314
EcoRI
ori
EcoRI
EcoRI
major
alternative
NcoI
AvaII
AvaII
EcoRI
EcoRI
(pBR322)
pSV2CAT
SV40
EARLY
POLY(A)
HpaI
SV40
EARLY
PROMOTER
CAT
HpaI
cleavage
3'
fill-in
(Klenow)
(pBR322)
SV40
EARLY
POLY(A)
SV40
EARLY
PROMOTER
CAT
DERIVATIVES pCP1 pCP2 pCP3 pCP4

Figure 3.13.

CAT assays using pSV2CAT derivatives. Plasmids pCP1, pCP2, pCP3, pCP4 and the parental vector pSV2CAT were transfected into Rat-1 cells, and extracts prepared 48 hours later were assayed for CAT activity. A negative control was provided by omitting DNA from the transfection mix (NO DNA). Each transfection was performed in duplicate, and autoradiographs of the chromatographically separated products are shown. The positions of chloramphenicol (CM) and of the 1- and 3-acetylated forms are indicated.


3-AcCM
1-AcCM
CM
pCP1
pCP2
pCP3
pCP4
pSV2CAT
NO DNA


SV40
EARLY
POLY(A)
SV40
EARLY
PROMOTER
CAT
DERIVATIVES
pCP1
pCP2
pCP3
pCP4

DISCUSSION

The sequence elements involved in eukaryotic transcription termination and 3' processing have been the subjects of a great deal of experimental attention in recent years. Termination of POLII transcription is likely to be a non-random event since, while some cellular genes many kilobases in length are transcribed in an equimolar manner along their entire length (Frayne *et al.*, 1984), the rate of transcription falls off within a few kilobases once the 3' end of a given gene has been reached (reviewed in chapter 1 and by Birnstiel *et al.*, 1985). Furthermore, there is evidence to suggest that, in some systems, differential termination of POLII transcription can be used in the regulation of gene expression (Mather *et al.*, 1984; Yuan and Tucker, 1984; Maderious and Chen-Kiang, 1984; Ben-Asher and Aloni, 1984; Hay and Aloni, 1984). Where the sequence elements involved in POLII termination have been studied in detail, they have been found to be rather complex and to encompass relatively large regions of DNA (Johnson *et al.*, 1986; Falck-Pedersen *et al.*, 1985) in comparison with the short sequence motifs involved in the regulation of transcript initiation, splicing and polyadenylation (see chapter 1). Direct study of termination events is further hampered because primary POLII transcripts appear to have highly heterogeneous 3' termini, and to be metabolically unstable *in vivo*. In at least one case, termination of transcription may depend upon prior cleavage of the nascent RNA at a poly(A) site (Falck-Pedersen *et al.*, 1985); conversely, there is evidence to suggest that termination can be used to influence the selection of a particular poly(A) site in a complex transcription unit (Mather et al., 1984; Yuan and Tucker, 1984).

Py-transformed cell lines exhibit marked differences in the pattern of polyadenylation of transcripts initiated within the integrated viral sequences (Fig. 3.1; Kamen *et al.*, 1979; Treisman, 1981). These differences probably result from differences in the cellular sequences adjacent to the site of viral integration; examination of Py transcription in transformed lines might be expected, therefore, to lead to the identification of cellular

sequences responsible for the observed differences in 3' RNA processing. During transformation, the sequences encompassing the Py major early poly(A) site are frequently detached from the active early transcription unit(s) (Lania *et al.*, 1981; Fried *et al.*, 1986: see appendix 2; see also chapter 4). Such 3'-truncated early regions can be regarded as a system for the selection of functional cellular poly(A) sites, which have been identified as the means by which efficient Py expression is often achieved after integration of the viral DNA (Ruley et al., 1982; Fenton and Basilico, 1981; B. Davies, personal communication). Somewhat analogous results have been obtained in experiments using Py genomes lacking a viral enhancer element; in two such cases viral expression was restored by cellular sequences with some of the properties expected of a cellular enhancer (Fried *et al.*, 1983; Ford *et al.*, 1985). In this study examination of cloned cellular sequences, derived from the 3' flank of the viral integration in the Py-transformed line 53-Rat, has led to the identification of a sequence element which has some of the properties expected of a POLII terminator.

Of the Py-transformed lines examined only two, namely 53-Rat and tsC3T3-2, displayed unusually high levels of polyadenylation at the alternative Py site (Fig. 3.1), while the total levels of Py RNAs were of the same order for all of the lines examined, regardless of their polyadenylation pattern. For 53-Rat, the majority of polyadenylated Py RNAs are cleaved at the alternative poly(A) site, as judged by northern blotting (Fig. 2.2) and S1 analysis (Figs 2.5, 3.1, 3.2). A small proportion of the Py transcripts in 53-Rat total cell RNA appear to extend beyond the viral-host junction (Fig. 3.3); at least some of these transcripts extend beyond an S1-hypersensitive site in the region of a (CT) poly(pyrimidine) sequence, between 456 and 494 nts downstream from the viral-host junction (Figs 3.4, 3.10). A low-level 3' end found only in the poly(A)$^+$ fraction corresponds to a site of polyadenylation 570 nts downstream from the viral-host junction, and 14 nts downstream from an AATAAA poly(A) site consensus sequence (Figs 3.4, 3.10). It will be interesting to discover whether or not this hexanucleotide and sequences

located immediately downstream are capable of directing polyadenylation at high efficiency when placed in a different sequence environment. If this were so, then the implication would be that the majority of Py nuclear transcripts in 53-Rat do not extend as far as this site, for one might otherwise expect such transcripts to be cleaved at this cellular poly(A) site rather than at the Py alternative site. The second of the AATAAA hexanucleotides found in the 53J sequence is followed 23 nts downstream by a TGTTG sequence; the first AATAAA element has no (T+G)-rich sequence between 14 and 24 nts downstream. The presence of a (T+G)-rich element in this region is correlated absolutely with the functionality of Py and SV40 AATAAA sequences (see chapter 1). In any event, no large quantity of steady-state nuclear or cytoplasmic RNA extends beyond the viral-host junction in 53-Rat (Figs 3.3, 3.4). No discrete poly(A)$^-$ 3' termini were detected; thus if nuclear transcripts do terminate within the region of 53-Rat sequence examined, this termination is highly heterogeneous and/or the poly(A)$^-$ RNAs are very unstable. Unstable RNAs with highly heterogeneous 3' termini appear to be the primary products of POLII termination in those genes where this phenomenon has been studied (Birnstiel *et al.*, 1985). The region of cellular DNA contained within the 53J clone did not appear to be transcribed in Rat-1 fibroblasts, as judged by northern analysis, but the possibility that these sequences are important for gene expression in other cell types cannot be ruled out. Since the majority of RNA-DNA duplexes containing the 53-Rat (CT) repeat were cleaved by S1, it would be interesting to use exonuclease VII, which does not cleave such structures (Johnson *et al.*, 1986), to investigate the low level nuclear transcripts which extend beyond this position.

DNA sequence determination showed that the high level of alternative poly(A) site use is not due to mutation of the site. This interpretation is supported by the finding that the central *Eco*RI fragment (Fig 3.2) can be re-circularised to produce an apparently wild-type A2 viral genome (Ruley *et al.*, 1987). Furthermore, any mutation of the 53-Rat sequence might have led to the generation of a novel hypersensitive site in S1 nuclease analysis with a probe derived from wild-type Py sequences; no such hypersensitive site

was detected (Fig. 3.3).

Cloned sequences from the region downstream from the 53-Rat Py insert are capable of directing the majority of polyadenylation to the alternative Py site in a plasmid reconstruction of the 53-Rat active transcription unit (p220.2, Fig. 3.6). In this reconstruction a known efficient poly(A) site, the Py major early site, was present downstream from the 53-Rat cellular sequence (53J, Fig.3.5). The fact that this site does not seem to be used in preference to the alternative site in p220.2 strongly suggests that the 53J sequences are at least partly responsible for the unusual pattern of polyadenylation observed in 53-Rat cells.

An increase in alternative site polyadenylation could be conferred by the 53J fragment if transcription were to terminate, or if primary transcripts were to be cleaved within the rat sequence. In either case, the capped primary transcript would contain the Py alternative site as the only site available for polyadenylation. A less extreme possibility is that the 53J sequences cause POLII to pause or proceed at a relatively slow rate; in this case the polyadenylation machinery would have a relatively increased probability of cleaving at the alternative site before traversal of a more efficient, downstream poly(A) site by the transcription complex. The 53J sequences may also act from a distance to increase the absolute frequency of cleavage at the alternative site; this interpretation is supported to a certain extent by the observation that the total levels of Py RNA in 53-Rat are of the same order as those in Py-transformed lines in which either the Py major poly(A) site or a nearby cellular poly(A) site is used by Py transcripts (Fig. 3.1). Increased efficiency of alternative poly(A) site cleavage would be substantiated if it could be shown that the rate of transcription initiation in this line is no higher than that observed in Py-transformed lines in which a downstream poly(A) site is used. Unfortunately, in attempts to investigate this point using the *in vitro* nuclear run-on technique, the level of Py transcription in 53-Rat cells appeared to be below the lower detection limit.

Reversal of the orientation of the 53J fragment resulted in the activation of a novel

splice site at Py nt. 1224 (Fig. 3.6). The corresponding 3' splice site was mapped to a position within the reversed 53J sequences; the novel splice sites are similar in sequence to the 5' and 3' Py splice sites described in earlier studies (Table 1.2). During the course of this work a report was published in which splicing from a position in the region of nt. 1224 to an acceptor in the Ad5 E1b gene was described (Davidson and Hassell, 1987). The E1b acceptor is no more similar to the 53J acceptor than the latter is to the known Py splice acceptors:

	intron	exon
p220.3 acceptor	TTTATCTACCTACAG	AGTCTGAT
Ad5E1b acceptor	AAGATACAGATTGAG	GTACTGAA

It would seem, then, that the nt. 1224 donor is specifically activated by the presence of a compatible 3' splice site, though any 3' site-specificity is not revealed by comparison of the available sequence data. Interestingly, transcripts from p220.3, pCFR2 and pCBgR2, which lack the Py alternative poly(A) site as a consequence of splicing between the novel splice junctions, are extended efficiently as far as the downstream, major Py site. This observation serves to suggest that sequences between the 53-Rat viral-host junction and a position 1152 nts downstream do not, when placed in the 'inverted' transcriptional orientation, prevent the extension of nuclear transcripts as far as the downstream, major poly(A) site. Note, however, that for plasmids where sub-fragments of these rat sequences are present in the same orientation, but the novel splice is not activated (pCAV2 and pCΔA2) use of the Py alternative site is favoured (Figs 3.8, 3.9). One may conclude that, when the alternative poly(A) site is not removed from the primary transcript by activation of the novel splice, the inverted 53J sequences are able to stimulate polyadenylation at this upstream position. There has been no report of the nt. 1224 donor being used during lytic Py infection; this may reflect the absence of any downstream

acceptor on monomeric Py pre-RNAs. Nevertheless, the formal possibility exists that the nt. 1224 donor could be paired with one of the previously described Py 3' splice sites following the generation of 'giant' multimeric Py transcripts. Such transcripts are believed to be formed at late times in Py infection (Kamen *et al.*, 1982). Splicing from nt. 1224 probably accounts for apparent 3' ends at this position detected at a low level in a number of Py-transformed lines (Kamen *et al.*, 1979).

The vector used for the initial plasmid reconstruction of the 53-Rat active transcription unit (p220.0, Fig. 3.5) was in some respects not ideally suited to further definition of the 53J transcriptional interference sequence. Firstly, the long end-labelled S1 probe used may be prone to artefactual results; this is because there is a finite possibility of the RNA-DNA duplex becoming nicked in a random manner during hybridisation and S1 digestion. The presence of a nick between the site of labelling and the position of a 3' end would decrease the observed signal corresponding to that 3' end; the relative yield of protected fragments would decrease with increasing fragment size. A second disadvantage of the p220.0 vector is the absence of a unique cloning site at the position of the 53-Rat viral-host junction. These shortcomings were avoided in pC253 (Fig. 3.7), the vector used in experiments designed to define more precisely the 53-Rat transcriptional interference element(s). In these experiments, the use of two end-labelled probes of similar size, one for each of the two Py poly(A) sites, served to increase the accuracy of estimation of the extent to which the downstream, major poly(A) site was used. In addition, only two probes were needed for comparison of transcription from derivatives carrying a variety of sub-fragments of the 53J sequence, since neither probe spanned the *Bam*HI site into which these fragments were cloned. The extent to which each fragment of rat DNA was able to stimulate alternative site polyadenylation was quantified after correction for the signals obtained by using pC253 alone.

The transcriptional interference sequences from 53-Rat do not appear to reside in a short, easily defined region (Figs 3.8, 3.9). Sub-fragments of the 1.4 kb rat sequence from 53J were all less active than the full-length fragment in promoting polyadenylation at

the upstream, alternative site. The effect is probably not, however, simply a function of the length of the fragment of rat DNA inserted between the two poly(A) sites; insertion into pC253 of a 1.0 kb fragment from the mouse immunoglobulin heavy chain gene gave a polyadenylation pattern indistinguishable from that of the vector. More likely, the promotion of alternative site use seen in 53-Rat, and in plasmids p220.2 and pCFR1, results from the combined effect of multiple, dispersed sequences within the 53J clone.

The sequence of the 53J rat DNA (Fig. 3.10) revealed no extensive homology with known genes. Nonetheless, specific comparison with previously defined POLII terminators from the mouse β_{major}-globin and sea urchin histone H2A genes (Citron *et al.*, 1984; Johnson *et al.*, 1986), revealed short stretches of homology. It remains to be seen whether or not these homologous regions are of importance for the transcriptional properties of the 53J sequence. The most extensive homology between the 53J and H2A sequences is centred on the (CT) dinucleotide repeat common to both; this sequence has, however, been shown to be unimportant for POLII termination in the H2A sequence (Johnson *et al.*, 1986).

In contrast with a potential POLII transcriptional terminator found downstream from the human gastrin gene (Sato *et al.*, 1986), the 53J sequence does not appear to act efficiently to terminate transcription when placed upstream from a poly(A) site (Fig. 3.13). The 53J sequence does, however, have some of the properties expected of a POLII terminator dependent for its function on the presence of an upstream poly(A) site; thus the majority of transcription through to the downstream poly(A) site may be prevented in the case of plasmids p220.2 and pCFR1 (Figs 3.6, 3.8). It will be interesting to discover if transcriptional read-through can be restored by disruption of the upstream, alternative poly(A) site. The 53J sequence also appears actively to promote polyadenylation at the upstream site; this interpretation is supported by the observation that the total polyadenylation signal observed for pCFR1 is of the same order as that observed for other pC253 derivatives (Fig. 3.9). Cleavage at the alternative site would thus seem to be

increased at the expense of major site cleavage in pCFR1 transcripts. *In vitro* transcription studies may prove useful in the further distinction between the possible functions of the 53J sequence. A complementary approach would be to introduce this sequence into an adenovirus vector such as that used by Falck-Pedersen *et al.* (1985). By using such a vector it should be possible to generate transcription levels high enough for *in vitro* nuclear run-on assays, which would be invaluable in detecting the POLII termination events which may underlie the 53-Rat polyadenylation pattern.

CHAPTER 4: RECOMBINATION EVENTS IN POLYOMA TRANSFORMATION

Results presented in the previous chapter show that the cellular sequences adjacent to the site of Py integration can, in certain instances, strongly influence the pattern of processing of transcripts initiated at the viral early promoter and extending beyond the viral-host junction. Viral and adjacent cellular sequences were molecularly cloned from three further lines with a view to investigating the transcriptional properties of these adjacent cellular sequences. The results of molecular cloning experiments are described in this chapter; preliminary transcriptional analyses are described in chapter 5.

Since there are no endogenous host sequences with extensive homology to Py, integration necessarily involves recombination between unrelated segments of DNA (illegitimate recombination). Results presented in this chapter extend previous findings relating to the mechanism of these recombination events. The sequences across seven previously unreported novel junctions are described; in no case are there more than three nucleotides of homology between parental strands at the site of the junction. In all three of the Py-transformed lines examined here, short stretches of Py DNA are located between an active, truncated early region and the adjacent cellular sequences. For the 82-Rat line, the arrangement of these 'filler' sequences strongly suggests that they were at one time contiguous with Py DNA found several kilobases away in the viral insert. A model for the events leading to the generation of this viral insert is presented.

OBJECTIVES

In the Py-transformed line 53-Rat, the sequences downstream from the viral-host junction influence the pattern of processing of transcripts initiated at the viral early promoter (see chapter 3). For 53-Rat, the effect of these downstream cellular sequences is to favour polyadenylation at the alternative viral poly(A) site; this site is also used at an elevated level in the Py-transformed mouse line tsC3T3-2 (Kamen *et al.*, 1979). The polyadenylation pattern observed in a second mouse line, tsC3T3-1, differs from that of tsC3T3-2 in that a smaller proportion of transcripts is cleaved at the alternative site (see chapter 5). These lines were derived, by single cell cloning, from a transformed line here termed tsC3T3U (M. Read, personal communication) which appeared to contain a single viral insertion. Preliminary restriction mapping data suggested that the viral inserts of tsC3T3-1 and tsC3T3-2 contained similar amounts of Py DNA, but differed slightly in the arrangement of these sequences (L. Lania, unpublished results). These findings suggested that rearrangement of the tsC3T3U insert has resulted in the generation of tsC3T3-1 and/or tsC3T3-2 (though the formal possibility remained that one of the single cell clones was derived from a small sub-population undetected in the uncloned line). It seemed possible, therefore, that the difference in polyadenylation pattern between tsC3T3-1 and tsC3T3-2 might be explained by differences in the 3' flanking DNA present in each line. In this study sequence analysis of novel recombinant junctions found in tsC3T3-1 and tsC3T3-2 was used to investigate the evolution in culture of the Py insert of these lines.

A different effect of downstream cellular sequences on Py gene expression is seen in the 82-Rat line. In this case Py-host fusion transcripts initiated at one of the three viral early promoters present fail to become polyadenylated efficiently at either a Py or a cellular poly(A) site. Cellular sequences adjacent to the affected 82-Rat early region were molecularly cloned with a view to examining this transcriptional phenomenon. Transcription data for this line, and for the tsC3T3 lines described above, are presented in

chapter 5.

During Py transformation, viral sequences become integrated into the host genome, allowing stable expression of the Py oncogene, middle T-Ag. Since at least several kilobases of cellular DNA are deleted at the site of integration (Stringer, 1982; Hayday *et al.*, 1982), a minimum of two recombination events - one at each end of the Py insert - are necessarily involved in the integration process. In all cases examined to date, there has been no extensive homology between the parental DNA strands at the site of recombination; the recombination events are therefore classed as 'illegitimate'. There are no obvious sequence specificities or structural features at the novel junctions. Nevertheless, one feature of many of these junctions is likely to be pertinent to the mechanism of their generation; this is the occurrence of a short (up to 5 nts) stretch of homology at the 'cross-over' site (Stringer, 1982; Ruley and Fried, 1983; Maruyama and Oda, 1984). Authentic integration of retroviral DNA has recently been demonstrated *in vitro* (Brown *et al.*, 1987). This approach will be of great value in studies of the enzymic steps required for proviral insertion; such steps are unlikely, however, to relate directly to Py integration, as the former (but not the latter) involves precise integration with no concomitant loss of cellular sequences (see chapter 1). Until an analogous *in vitro* system for the study of Py integration becomes available, the approach most likely to aid the elucidation of the mechanism of illegitimate recombination would seem to be the accumulation of a large database of Py/cell and Py/Py junction sequences generated *in vivo*. The majority of the results presented in this chapter relate to this point.

In the simplest cases, a partial Py monomer or tandem multimer is inserted at a single genomic site (Fried *et al.*, 1986: see appendix 2). In 82-Rat cells the arrangement of the Py insert is more complex; the viral sequences are interrupted by deletion/insertion events which result from clustered illegitimate recombination events (Ruley and Fried, 1983). In this study the sequence characterisation of the 82-Rat viral insert was completed, with a view to establishing a clearer model for the events which led to the generation of this

insert. Clustering of novel junctions has also been observed in the 7axB line (Hayday *et al.*, 1982). In this case a 37 nt. segment of cellular DNA has been transposed, in an inverted orientation, from a position 650 nts from the site of Py integration, to a position between the Py and adjacent cellular sequences (Williams and Fried, 1986). A further example is that of the 7axT line, which has a 12 bp segment, probably derived from Py early region sequences, interposed between the bulk of the Py insert and the adjacent cellular sequences (Hayday *et al.*, 1982). The occurrence of such short 'filler' DNA sequences within and adjacent to Py integrants is an intriguing feature of these presumed products of cellular recombination pathways.

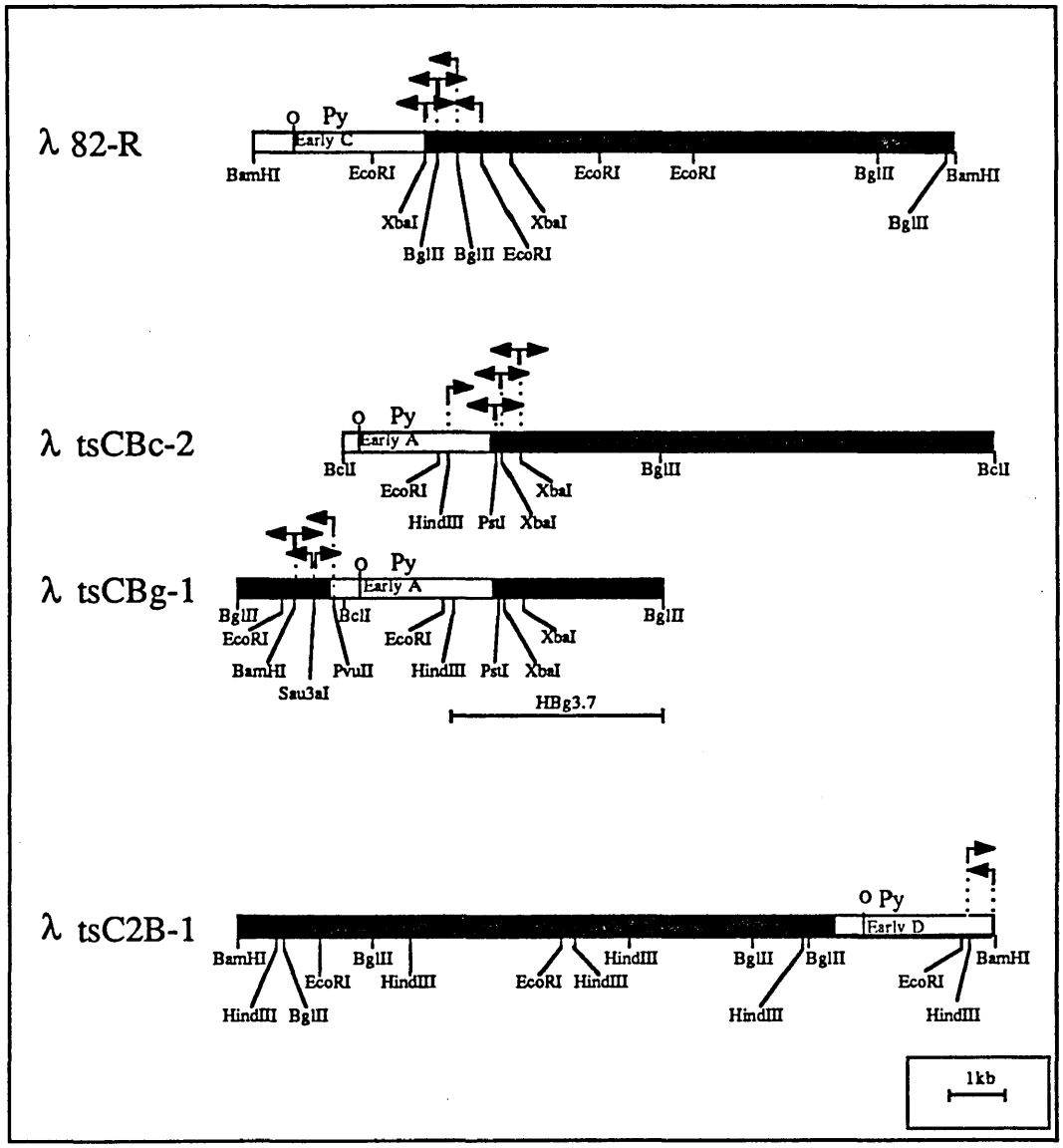
RESULTS

Molecular cloning of restriction fragments spanning viral-cell junctions.

After enrichment for Py-containing sequences by size fractionation on sucrose density gradients, bacteriophage λ libraries were created as described in the Materials and Methods section. For 82-Rat, a *Bam*HI library was constructed in the replacement vector L47.1 (Loenen and Brammar, 1980); the same vector was used in the construction of *Bcl*I and *Bgl*II libraries from tsC3T3-1. The replacement vector EMBL4 (Frischauf *et al.*, 1983) was used to generate a *Bam*HI library from tsC3T3-2. Approximately 10^5 recombinant phage from each of the four libraries were screened. Py-positive plaques were isolated from each library; partial restriction maps for the inserts of the four clones are presented in Fig. 4.1. The inserts of the *Bcl*I and *Bgl*II clones from tsC3T3-1 (λtsCBc-2 and λtsCBg-1 respectively) overlap by some 5 kb, and are apparently identical throughout this region. After sub-cloning into plasmid vectors of restriction fragments spanning the Py inserts, the detailed arrangement of viral sequences in each line was

Figure 4.1.

Restriction maps for the inserts of bacteriophage λ clones from 82-Rat, tsc3T3-1 and tsC3T3-2. For each clone Py sequences are shown as a white box containing the origin of replication (O). Flanking sequences of cellular origin are shown as black boxes. λ82R contains a 12 kb *Bam*HI fragment which spans the right-hand viral-host junction of 82-Rat (see Fig. 4.2). λtsCBc-2 (which contains a 10 kb *Bcl*I fragment) and λtsCBg-1 (which contains a 6.5 kb BglII fragment) together span the viral insert of tsC3T3-1. The location of the 3.7 kb *Hin*dIII-*Bgl*II fragment referred to in chapter 5 is indicated (HBg3.7). λtsC2B-1 contains Py sequences from tsC3T3-2 whose rightward limit is the *Bam*HI site shown, as well as ~11 kb of mouse sequences located to the left of the viral insert. Note that no *Bam*HI site is found in the early region of wild-type PyA2. For each clone the regions sequenced are denoted by horizontal arrows, the direction of sequencing being given by the direction of each arrow. The sequence information thus derived was used in the compilation of figures 4.2 to 4.5.


λ 82-R
o Py
Early C
BamHI
EcoRI
XbaI
BglII
BglII
EcoRI
XbaI
EcoRI
EcoRI
BglII
BamHI
BglII
λ tsCBc-2
o Py
Early A
BclI
EcoRI
HindIII
PstI
XbaI
BglII
BclI
λ tsCBg-1
o Py
Early A
BglII
EcoRI
BamHI
Sau3aI
PvuII
BclI
EcoRI
HindIII
PstI
XbaI
BglII
HBg3.7
λ tsC2B-1
o Py
Early D
BamHI
HindIII
EcoRI
BglII
HindIII
BglII
EcoRI
HindIII
HindIII
BglII
HindIII
BglII
EcoRI
HindIII
BamHI
1kb

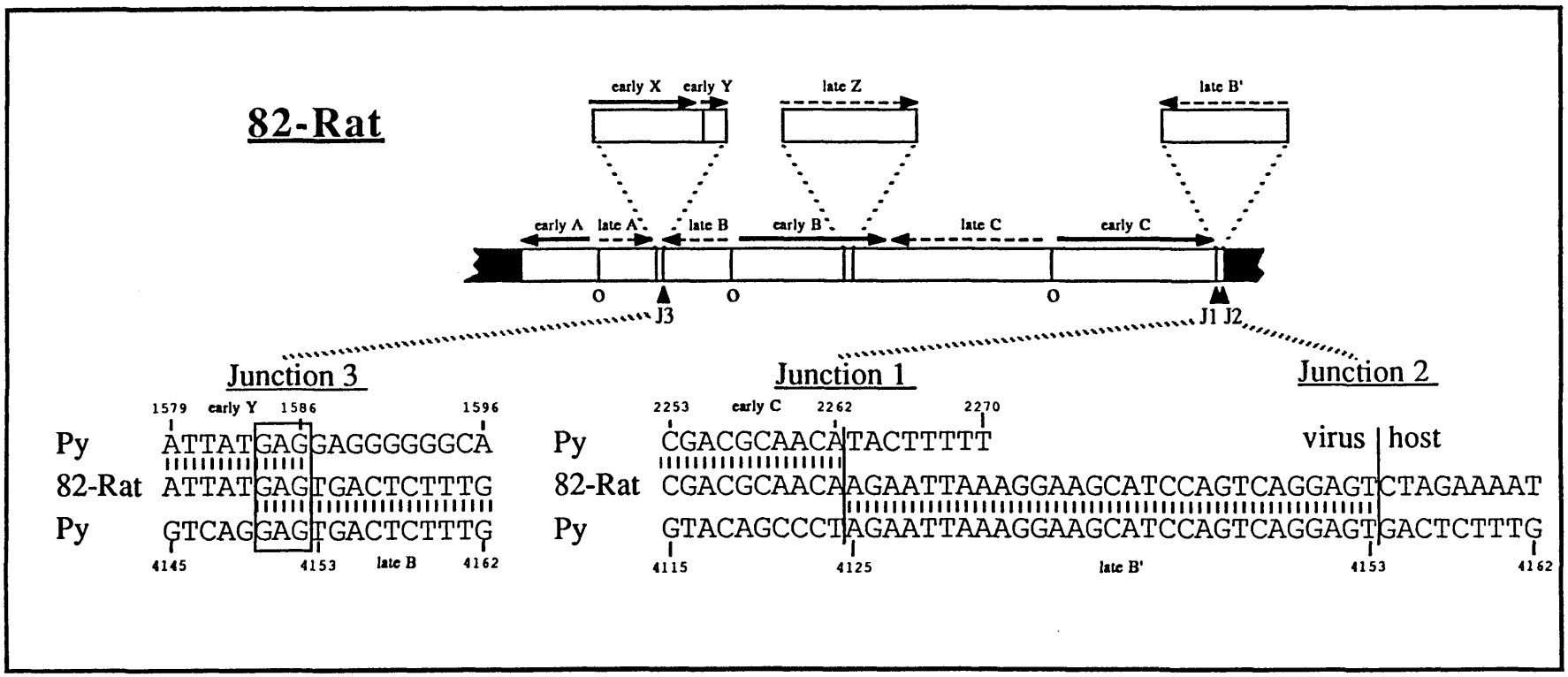
determined, initially by restriction mapping and subsequently by sequence analysis. Preliminary restriction mapping of the tsC3T3-2 insert by Southern analysis (L. Lania, unpublished results) had suggested that the orientation of the Py sequences was such that the *Bam*HI fragment cloned as λtsC2B-1 should contain several kilobases of mouse DNA downstream from the viral early region. In fact, this interpretation of the Southern data had been incorrect and no mouse 3' flanking sequences were cloned (Fig. 4.1). When this had been discovered, an *Eco*RI library was constructed from size-selected (9-10 kb) tsC3T3-2 cell DNA in λgt-10, but despite the fact that some 1.5 x 10^5 recombinants were screened, no Py-positive plaques spanning the right-hand tsC3T3-2 junction were isolated. This may indicate that sequences downstream from the tsC3T3-2 insert are inhibitory to phage growth.

Further clustered illegitimate recombination events in the 82-Rat Py insert.

A previous study of the 82-Rat Py-transformed line identified two deletion/insertion structures which are the result of clustered illegitimate recombination events within the body of the viral integrant (Ruley and Fried, 1983). These short Py sequences (partial early regions X and Y and partial late region Z) and the sites of deletion/insertion are shown in Fig. 4.2. The novel junction at the left-hand extremity of the Py sequences had also been shown to involve straightforward illegitimate recombination, with an abrupt switch from Py to cellular sequences (Ruley *et al.*, 1987). In contrast sequence analysis of λ82-R subclones, using the strategy summarised in Fig. 4.1, revealed that a short (~28 bp) 'filler' segment of Py late region DNA was present at the right-hand extremity of the viral integration (late B', Fig. 4.2). This short segment of late region DNA had not been identified by previous Southern analyses. Since no further Py-containing sequences were detected in the region of λ82-R lying to the right of late B' (Fig. 4.2), the map presented here represents a complete description of the 82-Rat viral integrant.

Figure 4.2.

The arrangement of recombinant joins in 82-Rat. The map in the upper part of the figure shows the Py sequences (white) at the site of integration into host DNA (black) in 82-Rat. The left-hand half of the map is reproduced from Ruley and Fried (1983), whereas the right-hand half is derived from the results of this study. Short segments of early region denoted X and Y and of late region denoted Z and B' (not drawn to scale) are present at the three sites shown. The positions of newly-discovered junctions 1 and 2, and the position of the junction (3) between early segment Y and late region B previously reported (Ruley and Fried, 1983) are indicated. Sequence information for novel junctions in 82-Rat is given in the lower part of the figure. For junctions 1 and 3 (J1, J3) the central line of sequence represents that found in 82-Rat, with the lines above and below representing the two parental strands. For junction 2 (J2) only one parental sequence, that of Py (shown in the lower line) is currently known. Homology between the donor strands and the 82-Rat sequence is denoted by vertical dashes and numbers refer to Py nucleotide positions. For junction 1, the cross-over site can be determined precisely from the three sequences: the position of this site is indicated by the vertical bar. For junctions 2 and 3, there is a degree of ambiguity concerning the precise position of the cross-over, and for junction 3 this is indicated by a box. For junction 2, the extent of this ambiguity is not yet known, since cellular sequences from the 'unoccupied' integration site have not been molecularly cloned.


82-Rat
early X
early Y
late Z
late B'
early A
late A
late B
early B
late C
early C
J3
J1 J2
Junction 3
Junction 1
Junction 2
1579
early Y
1586
1596
Py ATTATGAGGAGGGGGGCA
82-Rat ATTATGAGTGACTCTTTG
Py GTCAGGAGTGACTCTTTG
4145
4153
late B
4162
2253
early C
2262
2270
Py CGACGCAACATACTTTTT
virus | host
82-Rat CGACGCAACAAGAATTAAAGGAAGCATCCAGTCAGGAGTCTAGAAAAT
Py GTACAGCCCTAGAATTAAAGGAAGCATCCAGTCAGGAGTGACTCTTTG
4115
4125
late B'
4153
4162

The sequences at the novel junctions between 82-Rat early region C and the 'filler' late B' (junction 1), and between late B' and the adjacent cellular DNA (junction 2) are presented in Fig. 4.2. For the Py/Py junction (1) the sequences of both parental strands are known, and the precise site of recombination can be seen to lie, for one strand, between Py nts 2262 and 2263 and, for the other, between Py nts 4124 and 4125. At present, no sequence information is available for the displaced rat DNA at the 'unoccupied' site of the 82-Rat Py insertion. As a consequence, only one of the two parental strands which contributed to junction 2 (between the filler late B' and adjacent rat DNA) is of known sequence, and the precise cross-over site at this junction cannot be deduced. By analogy with known illegitimate recombinant junctions, one might expect there to be, on average, 2 bp of homology at the cross-over site. In this case the site would lie between Py nts 4151 and 4154. This finding is intriguing in the light of sequence information described by Ruley and Fried (1983) for the junction between partial early region Y and late region B (junction 3, Fig. 4.2). For the late strand, this recombination event took place between Py nts 4149 and 4153; the ambiguity in this case results from the 3 nts of homology at the cross-over site. Despite this ambiguity, the probability of finding two of the eight 82-Rat novel junctions in this short region of Py sequence is small (of the order of 0.004), assuming random recombination within the Py genome.

The organisation of Py sequences in tsC3T3-1 and tsC3T3-2.

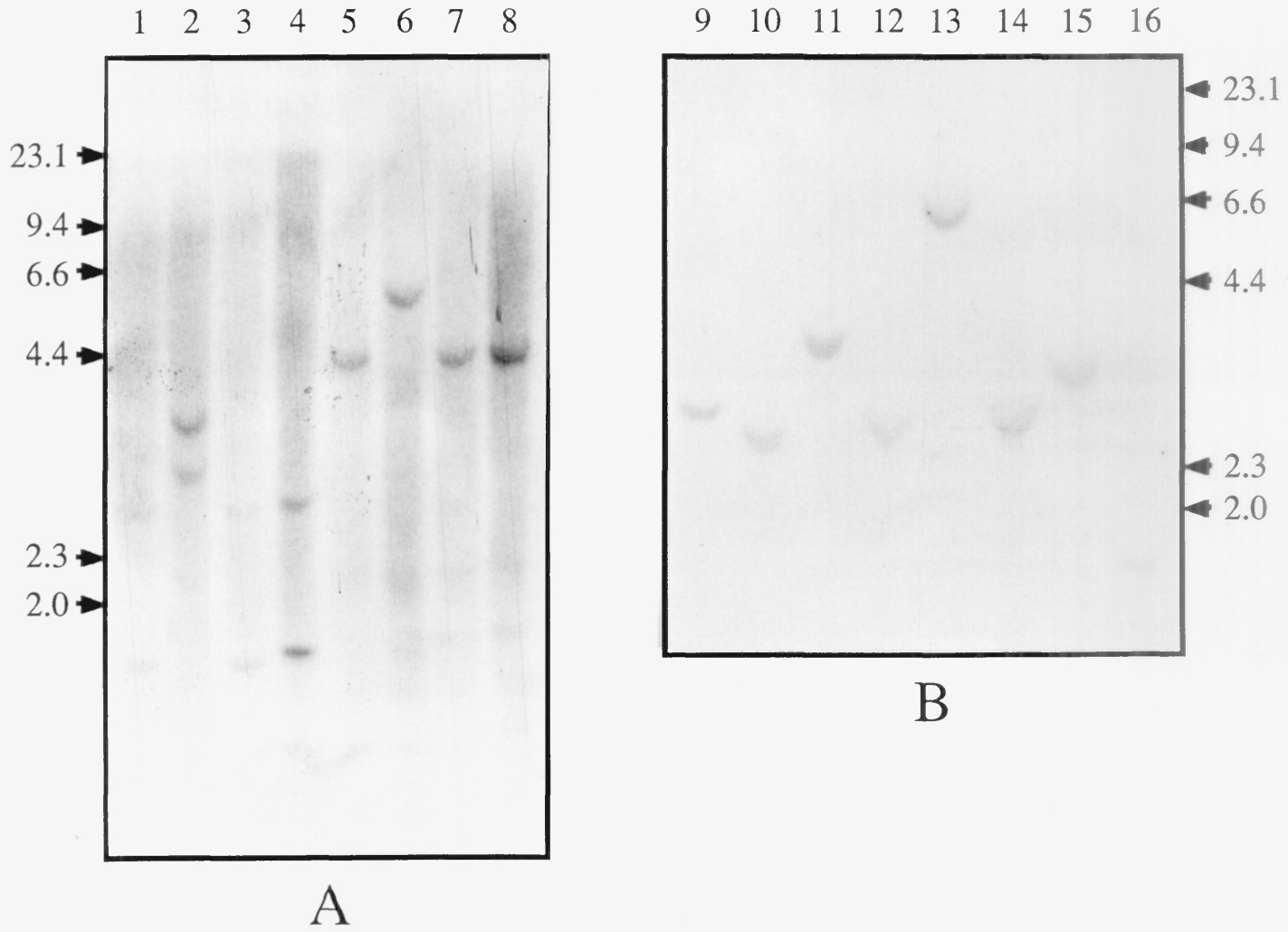
The extent of the difference between the single-cell clones tsC3T3-1 and-2 was initially assessed by Southern blotting (Fig. 4.3). Far from being broadly similar, as suggested by an earlier study (L. Lania, unpublished results), the two viral inserts appeared to differ markedly, as evidenced by differences in the *Hin*dIII and *Bgl*II restriction patterns obtained in Southern analysis using a Py probe (Fig. 4.3A), and in the

Figure 4.3.

Southern analyses of tsC3T3 cell lines. Total cell DNA (20μg) from tsC3T3U (lanes 1 and 5), tsC3T3-1 (lanes 2, 6 and 9-16), tsC3T3-2 (lanes 3 and 7) and tsC3T3-3 (lanes 4 and 8) was digested with *Hin*dIII (lanes 1-4 and 11), *Bgl*II (lanes 5-8 and 13), *Eco*RI (lane 9), *Eco*RI and *Bam*HI (lane 10), *Hin*dIII and *Bam*HI (lane 12), *Bgl*II and *Eco*RI (lane 14), *Bgl*II and *Hin*dIII (lane 15) and *Pst*I (lane 16). After Southern transfer, filters were probed with either a *Bam*HI linear clone of Py in pAT153 (lanes 1-8) or with a *Bam*HI-*Eco*RI fragment spanning the promoter-proximal part of the Py early region (lanes 9-16).

*Hin*dIII-restricted DNA from tsC3T3U, -2 and -3 contains Py-containing bands of 1.8 and 2.5 kb (lanes 1, 3 and 4), whereas DNA from tsC3T3-1 gives bands of 2.8 and 3.4 kb (lane 2). Similarly, whereas the Py-containing fragment in a *Bgl*II digest is one of 4.2 kb for tsC3T3U, -2 and -3 (lanes 5, 7 and 8), the fragment from tsC3T3-1 is 6.5 kb in length (lanes 6 and 13).

Further analysis of the tsC3T3-1 insert (panel B) shows that the restriction maps of the two overlapping λ clones derived from this line (Fig. 4.1) correspond to the map of the Py insert and surrounding cellular sequences. The use of the Py *Bam*HI-*Eco*RI probe results in the detection of only one of the Py-containing fragments in lanes 9-12, 14 and 15, and this facilitates the construction of a partial restriction map. The 2.4 kb *Eco*RI fragment (lane 9) is internal to the 6.5 kb *Bgl*II fragment (lane 13) and is thus not cleaved by *Bgl*II (lane 14). The *Bam*HI site is located within the *Eco*RI fragment, some 200 bp from one end (lane 10) and 2.3 kb from the Py *Hin*dIII site (lane 12); the latter site is in turn 2.8 kb from the (left-hand) *Bgl*II site. All of these data are in agreement with the tsC3T3-1 map presented in Fig. 4.1.


1 2 3 4 5 6 7 8
23.1
9.4
6.6
4.4
2.3
2.0
A
9 10 11 12 13 14 15 16
23.1
9.4
6.6
4.4
2.3
2.0
B

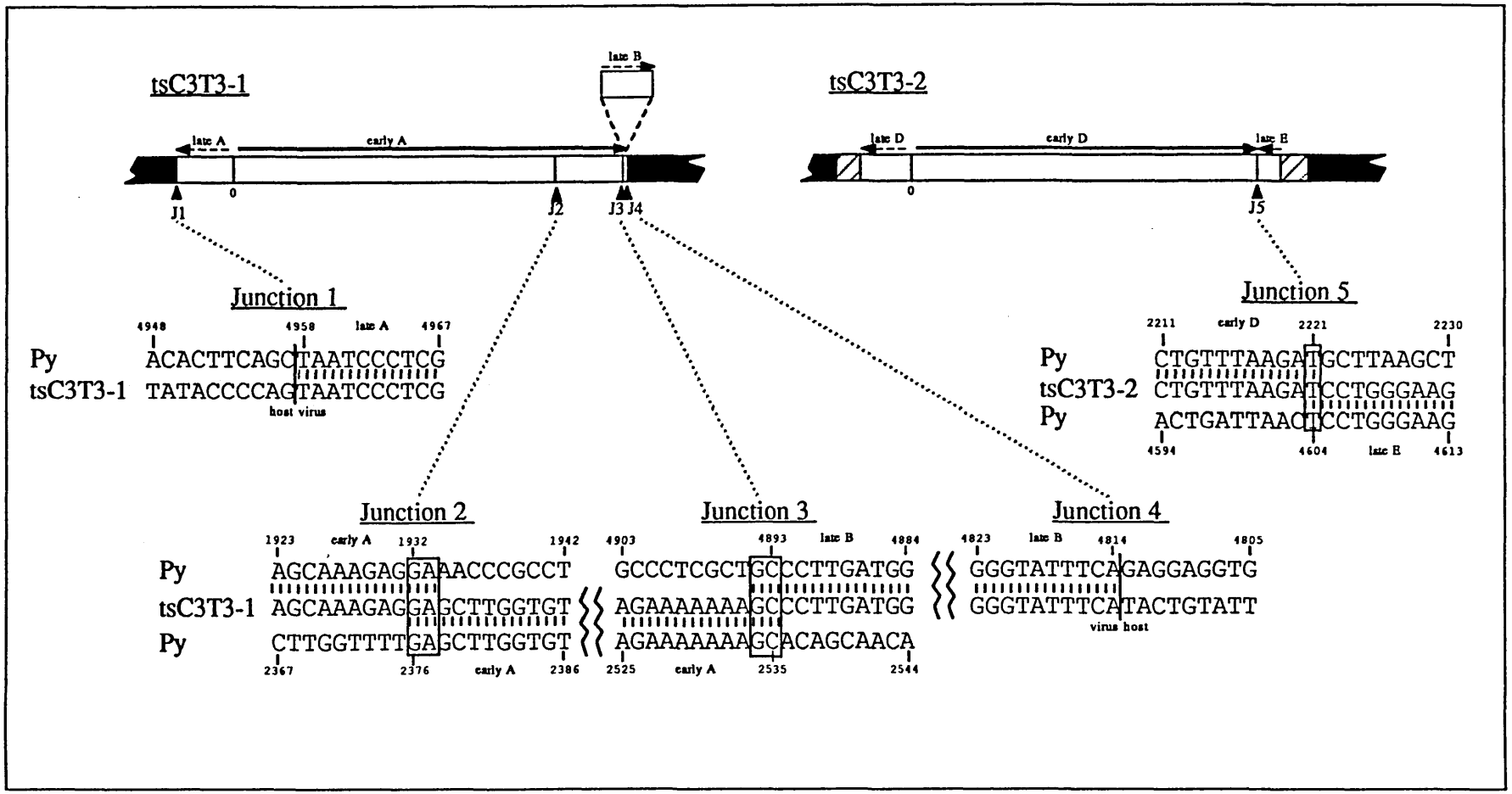
detailed restriction maps of λ clones (Fig. 4.1). Restriction mapping and sequence analysis of the Py and adjacent mouse DNA from λtsCBc-2, λtsCBg-1 and λtsC2B-1 made possible a complete description of the Py insert of tsC3T3-1, as well as a partial description of the insert of tsC3T3-2 (Fig. 4.4). For tsC3T3-1, the left-hand viral-host junction (Fig. 4.4, junction 1) is a straightforward, abrupt transition from mouse DNA to the partial Py late region A. The viral sequences of the single early region (A) are interrupted by an apparent internal deletion (Fig. 4.4, junction 2) within the sequences unique to the large T-Ag coding frame (see also Fig. 2.1). At the right-hand end of the insert, a short (~79 bp) filler of Py late region DNA (Fig. 4.4, late B) is interposed between early region A and the adjacent mouse sequences (Fig. 4.4, junctions 3 and 4). The arrangement of viral and adjacent cellular sequences from tsC3T3-1 deduced from Southern blotting data (Fig.4.3B) corresponds to that deduced from restriction mapping of λ clones (Fig. 4.1); furthermore, the two λ clones which span the tsC3T3-1 viral insert (Fig. 4.1) overlap by some 5.5 kb, and are apparently identical throughout the overlap. One can therefore be confident that the arrangement of tsC3T3-1 sequences shown in Fig. 4.4 is that which exists in the cell line.

The Py insert of tsC3T3-2 is similar to that of tsC3T3-1 in that the majority of the viral sequences are present as single early and late regions (D, Fig. 4.4). Again, a fragment of Py late region DNA (late E) is found between early region D (Fig. 4.4, junction 5) and the adjacent mouse sequences. A minimum estimate for the length of this fragment is 28 bp, this being the distance between junction 5 (Fig. 4.4) and the Py *Bam*HI site which marks the rightward extent of sequences cloned in λtsC2B-1 (Fig. 4.1). Note that the late region segments B and E of tsC3T3-1 and tsC3T3-2 are present in opposite orientations.

The sequences at all of the tsC3T3-1 novel junctions, and at the junction between early region D and late region E of tsC3T3-2, are presented in the lower part of Fig. 4.4. For junctions involving two Py DNA segments, the extent of the homology between parental strands at the cross-over site can be determined, since both parental sequences are known.

Figure 4.4.

The arrangement of recombinant joins in tsC3T3-1 and tsC3T3-2. The maps in the upper part of the figure show Py sequences (white) at the sites of integration into mouse DNA (black) in the transformed lines tsC3T3-1 and tsC3T3-2. The location of each Py origin of replication (o) and partial early and late region (horizontal arrows) are shown. For tsC3T3-2, the precise positions of the left- and right-hand viral-host junctions are not yet known, and this is denoted by hatched boxes. The positions of junctions 1-5 are indicated (J1-J5); the sequences of these junctions are presented in the lower part of the figure. For junctions 2, 3 and 5 the central line of sequence represents that found in the transformed line, with the lines above and below representing the two parental strands. For junctions 1 and 4 only the Py parental sequence, shown in the upper line, is currently known. Homology between the donor strands and the sequence found in the transformed line is denoted by vertical dashes, and numbers refer to Py nucleotide positions. For all of the junctions there is a degree of ambiguity concerning the precise position of the cross-over, and this is indicated by boxes. For junctions 1 and 4, the extent of this ambiguity is not yet known, since cellular sequences from the 'unoccupied' integration site have not been molecularly cloned.


tsC3T3-1
late B
late A
early A
0
J1
J2
J3 J4
tsC3T3-2
late D
early D
late E
0
J5
Junction 1
4948
4958
late A
4967
Py ACACTTCAGCTAATCCCTCG
tsC3T3-1 TATACCCCAGTAATCCCTCG
host virus
Junction 5
2211
early D
2221
2230
Py CTGTTTAAGATGCTTAAGCT
tsC3T3-2 CTGTTTAAGATCCTGGGAAG
Py ACTGATTAACTCCTGGGAAG
4594
4604
late E
4613
Junction 2
Junction 3
Junction 4
1923
early A
1932
1942
4903
4893
late B
4884
4823
late B
4814
4805
Py AGCAAAGAGGAAACCCGCCT GCCCTCGCTGCCCTTGATGG GGGTATTTCAGAGGAGGTG
tsC3T3-1 AGCAAAGAGGAGCTTGGTGT AGAAAAAAAGCCCTTGATGG GGGTATTTCATACTGTATT
Py CTTGGTTTTGAGCTTGGTGT AGAAAAAAAGCACAGCAACA
virus host
2367
2376
early A
2386
2525
early A
2535
2544

Thus junctions 2 and 3 from tsC3T3-1 both have 2 nts of cross-over site homology, while junction 5 from tsC3T3-2 has 1 nt. homology. For the viral-host junctions of tsC3T3-1 (Fig. 4.4, junctions 1 and 4) the parental mouse sequences are unknown, as are the extents of cross-over site homology.

DISCUSSION

The cellular processes which mediate the joining of DNA strands of unrelated sequence (illegitimate recombination) are likely to be complex, as the recombinant products fall into a number of structural classes (Roth and Wilson, 1986). Results presented here reveal the arrangement of viral sequences and the structures of novel recombinant junctions in three Py-transformed cell lines. These results must be accommodated into any model for the mechanism(s) of illegitimate recombination.

The recombination cross-over sites described (Figs 4.2, 4.4) are typical examples of the structures found at novel junctions both in Py (or SV40) transformed cells and in cells containing DNA deletions, translocations or amplifications. No common structural or sequence features were found at the sites of recombination. The existence of up to 3 nts of homology between parental strands at the cross-over site is in line with the template-directed and post-repair models for illegitimate recombination proposed by Hasson *et al.* (1984) and by Roth and Wilson (1986). Junctions which show no inter-strand homology at the cross-over site, such as the 82-Rat join between early region C and the late B' filler (Fig. 4.2, junction 1), may be products of an alternative recombination mechanism involving the ligation of single-stranded DNAs (Roth and Wilson, 1986). One possible candidate for the enzyme which endonucleolytically cleaves the parental DNA strands prior to recombination is topoisomerase I, and an association of SV40 recombinant junctions with topoisomerase I *in vitro* cleavage sites has been reported (Bullock *et al.*, 1985). Of the junctions described here, only the tsC3T3-2 internal junction (Fig. 4.4,

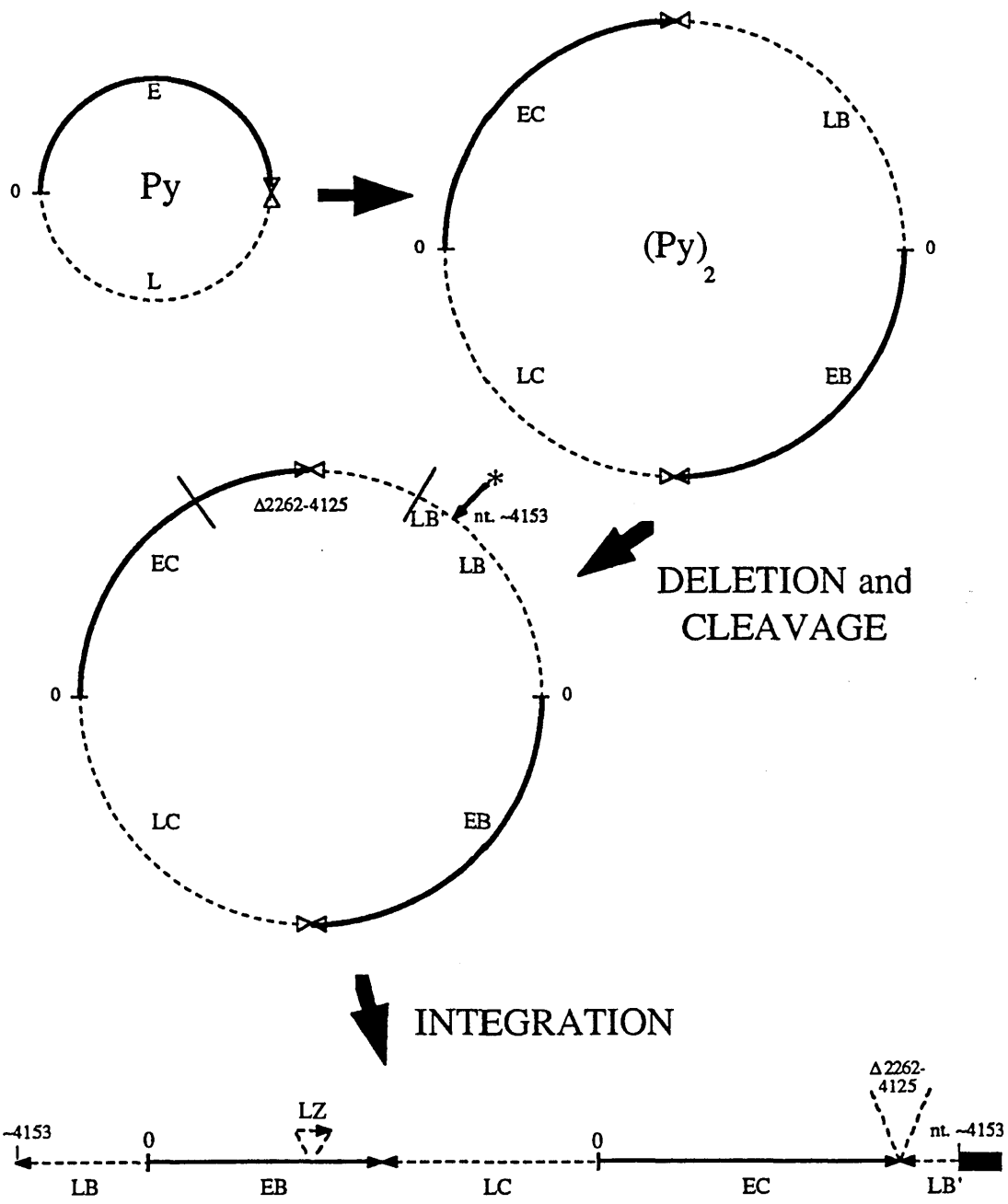
junction 5) conforms to the consensus cleavage site for eukaryotic topoisomerase I, $^{5'}(^{A}/_{T})(^{G}/_{C})(^{A}/_{T})$T(cut), derived by Been *et al.* (1984). In a random sequence, one would expect to find this consensus once in every 32 nts; its significance is therefore questionable. Using a large computer database of recombinant junctions, an attempt is being made to discover any sequence specificity at the cross-over site (P. Bullock, personal communication). The sequences described here have been made available to this statistical analysis.

The arrangement of viral DNA within the Py insert of the transformed line 82-Rat has previously been shown to be complex, though restricted to a single chromosomal location (Ruley and Fried, 1983). Results presented here revealed further complexity in this viral insert; specifically, a ~28 bp segment of Py late region DNA (Fig. 4.2, late B') is located between the previously described early region C (Fig. 4.2; Ruley and Fried, 1983) and adjacent rat DNA. Short stretches of filler DNA have been found to be interposed between two longer parental sequences at a number of recombinant junctions (Williams and Fried, 1986 and references therein). In some instances the filler sequences may be derived by transposition from a distant chromosomal site (Williams and Fried, 1986) but, in the case of the 82-Rat late B' DNA, it appears that the filler was attached to the viral sequences of late B (Fig. 4.2) before the stable Py integrant was established. This conclusion is based on the finding that Py late region sequences at the junction between late B' and flanking rat DNA (Fig. 4.2, junction 2) are contiguous with the Py late region sequences at the junction between late B and early Y (Fig. 4.2, junction 3). Since the arrangement of Py sequences shown in Fig. 4.2 was detected soon (~30 generations) after transformation (Ruley and Fried, 1983), it seems probable that this arrangement is identical to that of the initial integrant. Late B' and late B (Fig. 4.2) probably, therefore, were separated before or during the initial integration, rather than during subsequent culture *in vitro*.

A model for the generation of the right-hand side of the 82-Rat insert is presented in Fig. 4.5. Early and late regions B and C (Fig. 4.2) could be derived from a dimeric Py

Figure 4.5.

A model which accounts, in part, for the arrangement of viral sequences in 82-Rat. A single Py genome is represented in the upper left; E and L denote the early and late regions, which are also drawn as solid and dashed arrows, respectively. The Py origin of replication is also indicated (O). A Py dimer (upper right) could be generated by a replication error or by homologous recombination between two monomers. For ease of comparison with the map of 82-Rat (Fig. 4.2), the two early and late regions are labelled B and C. The dimeric molecule suffers an internal deletion of early region C (EC) and late region B (LB) sequences between nts 2262 and 4125 (Δ2262-4125), and is also cleaved endonucleolytically within late region B, in the region of Py nt. 4153; the site of cleavage is indicated by an asterisk. The linearised molecule then recombines at the right-hand end with genomic rat DNA (thick black line). The insertion/deletion event leading to acquisition of late region fragment Z (LZ) could have occurred either before or after integration. The derivation of partial early regions A, X and Y and of late region A (Fig. 4.2) has been discussed previously (Ruley and Fried, 1983).


E
Py
L
0
EC
LB
(Py)2
LC
EB
Δ2262-4125
nt. ~4153
DELETION and
CLEAVAGE
INTEGRATION
LZ
nt. ~4153
LB
EB
LC
EC
LB'

molecule which suffered an internal deletion of sequences between nts 2262 and 4125 as well as an endonucleolytic cleavage at nt. ~4153 within late region B. Integration of the resultant linear molecule into the rat genome would result in the observed arrangement of viral DNA (Figs 4.2, 4.5). If a free end at Py nt. ~4153 was one donor strand for the right-hand novel junction (Fig. 4.2, junction 2), then no Py sequences were lost upon recombination. This is in marked contrast with the loss of at least several kilobase pairs of cellular DNA observed at the site of viral integration in Py- and SV40-transformed cells (Hayday *et al.*, 1982; Stringer, 1982). It is, however, conceivable that an end-to-end ligation of Py and rat sequences generated this right-hand junction prior to integration of the recombinant structure into the chromosome. Partial early regions A, X and Y, and late regions A and Z (Fig. 4.2) may have been generated by independent events, and subsequently ligated with the Py sequences depicted in Fig. 4.5 either before or during integration. The observation of eight novel recombinant junctions within the region of the 82-Rat viral insert suggests the possibility that this cell line, or the region of the genome into which the Py DNA is integrated, is unusually susceptible to illegitimate recombination. This possibility can be investigated by examination of the 'unoccupied' integration site in the parental cell line, Rat-1; the cellular sequences adjacent to the 82-Rat insert may be rearranged with respect to this unoccupied site.

Further examples of viral filler sequences were found in the Py-transformed mouse lines tsC3T3-1 and tsC3T3-2 (Fig. 4.4). For tsC3T3-1, this filler is ~79 bp in length; the length of the filler sequence in tsC3T3-2 is at least 28 bp, but its precise size is not yet known. The tsC3T3-1 Py insert is further complicated by the presence of an internal deletion of early region sequences between Py nts ~1932 and ~2376.

Southern and transcriptional analyses suggested that the arrangement of Py sequences in tsC3T3-2 is identical to that found in the transformed line (tsC3T3U) from which it was cloned (Figs 4.3, 5.2). The tsC3T3-1 line, derived from the same line, appears to differ in its arrangement of Py sequences from tsC3T3U, and hence from tsC3T3-2. The most

likely explanation for this difference was initially thought to be that the tsC3T3-1 Py insert had evolved from the tsC3T3-2-type insert during culture *in vitro*. Changes in the state of the integrated viral sequences have been reported both for Py- and for SV40-transformed lines (Birg et al., 1979; Hiscott et al., 1981; Bender and Brockman, 1981). Though similar amounts of Py DNA are present in the two tsC3T3 cloned lines, the restriction maps of the surrounding cellular sequences are entirely different (Figs 4.1, 4.3). Furthermore, tsC3T3-1 contains sequences between Py nts 2376 and 2535 which are absent from the tsC3T3-2 insert (Fig. 4.4). Thus it seems likely that tsC3T3-1 contains a Py integration pattern derived from a small sub-population of the original tsC3T3U line undetected by Southern analysis.

CHAPTER 5: POLYOMA TRANSCRIPT PROCESSING IN THE 82-RAT AND tsC-3T3 TRANSFORMED LINES

Many Py-transformed lines contain truncated viral early regions whose primary transcripts can differ markedly in their processing patterns. The 82-Rat line contains three partial early regions, two of which give rise to primary viral-host fusion transcripts. While the RNAs derived from one of these early regions are efficiently polyadenylated within cellular sequences and exported to the cytoplasm, RNAs derived from the other fail to become polyadenylated efficiently at a cellular or viral site. In contrast to the 53-Rat transcription pattern described in chapter 3, the alternative viral poly(A) site does not become activated in any of the 82-Rat early transcripts.

In addition to 53-Rat, tsC3T3-2 is a second Py-transformed line in which the alternative poly(A) site is used at a relatively high level. The polyadenylation patterns and nuclear-cytoplasmic distributions were examined for Py RNAs from this line and from the apparently related lines tsC3T3U, tsC3T3-1 and tsC3T3-3. The transcription pattern of tsC3T3-1 was found to differ from that of the three other lines in that a less elevated level of alternative site use was observed.

Differences in Py RNA processing between the cell lines described here would appear to result from the different cellular sequences adjacent to each truncated viral early region. Preliminary results using plasmid reconstructions of the 82-Rat and tsC3T3-1 early regions suggest that these transcriptional phenomena may be reproduced only to a limited extent using this approach.

OBJECTIVES

Truncated Py early regions lacking the major early poly(A) site are a common feature of Py-transformed cells (Fried *et al.*, 1986: see appendix 2). These partial transcription units can be regarded as tools with which a variety of cellular transcriptional processes can be explored. In the 53-Rat line, the cellular sequences adjacent to the viral insert act to increase the level of cleavage at the alternative Py poly(A) site; in other lines a cellular poly(A) site downstream from the viral insert is used (see chapters 1 and 3). A third pattern of processing is found in the 82-Rat line, the Py insert of which contains three partial early regions (early A, B and C, Fig. 4.2). An earlier study (Treisman, 1981) had suggested that transcripts from partial early region C were not exported efficiently to the cytoplasm, but were instead accumulated in the nuclei of 82-Rat cells. In contrast, the hybrid viral-host transcripts of early region A were found to be exported into the cytoplasm with high efficiency (Treisman, 1981), after polyadenylation at two cellular sites in the 3' flanking sequences (B. Davies, personal communication). The 82-Rat early regions A and C, which contain identical 5' expression sequences and are integrated into the same region of the rat genome, thus display very different patterns of RNA processing. A plasmid reconstruction approach had been used to characterise transcriptional properties of sequences from 53-Rat (see chapter 3). It was reasoned that such an approach might be used to investigate the extent to which the lack of polyadenylation of 82-Rat early region C transcripts results from the nature of cellular sequences downstream from this partial early region.

Single cell clones tsC3T3-1, -2 and -3 were derived from the uncloned Py-transformed line tsC3T3U (see also chapter 4). A previous study had shown that the pattern of polyadenylation of Py RNAs in tsC3T3-2 is very similar to that of 53-Rat, with the majority of Py mRNAs being cleaved at the alternative poly(A) site (Kamen *et al.*, 1979; see also chapter 3). In this chapter the polyadenylation patterns of tsC3T3U, -1 and

-3 are described. Differences in the Py polyadenylation patterns of these lines might be expected to result from any alterations in the structure of the Py insert during growth in culture. It was initially hoped that identification of such polyadenylation differences and structural alterations might lead to a description of the sequence elements involved. As with 53-Rat and 82-Rat, plasmid reconstructions were to be used to assess the transcriptional properties of cloned flanking sequences from the tsC3T3 lines.

RESULTS

Processing of Py RNAs derived from the three 82-Rat partial early regions.

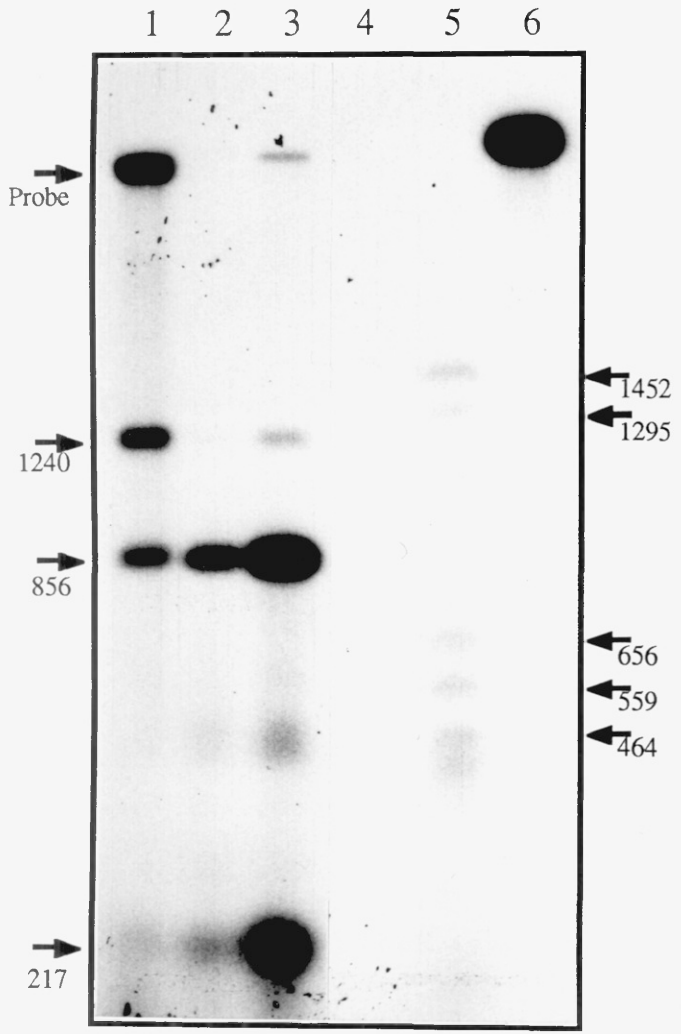
In an earlier study of 82-Rat transcription, the nuclear and cytoplasmic RNAs derived from the three partial Py early regions had been examined in separate experiments (Treisman, 1981). One conclusion drawn from these experiments was that early region C transcripts accumulate at high levels in the nucleus. This would suggest that sequences downstream from early region C are able to stabilise the hybrid viral-host nuclear RNAs. Stable nuclear RNAs are not frequently observed products of POLII transcription, and the 82-Rat cellular sequences involved were therefore deemed worthy of further investigation.

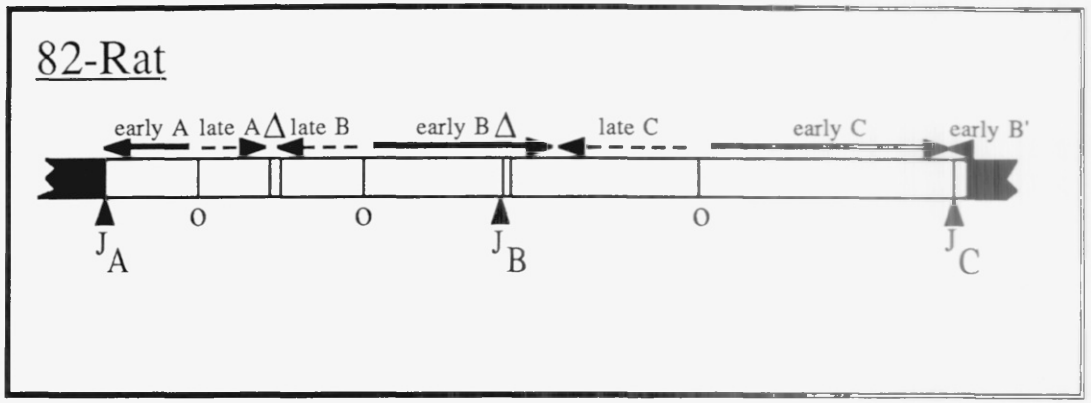
A 3' S1 nuclease analysis was performed with an *Ava*I probe labelled at Py nt. 1022, in such a way that the nuclear and cytoplasmic RNAs originating from 82-Rat early regions A, B and C could be examined simultaneously (Fig. 5.1); this analysis gave rise to a number of points of interest. For each partial early region, the distribution of transcripts between the nuclear and cytoplasmic compartments is quite distinctive (compare lanes 1 and 3, Fig. 5.1). Of the three partial early regions, early region B most closely resembles the wild-type Py sequence, from which it differs only in that it contains a short deletion/insertion structure (late Z, Fig. 4.2). The majority of early region B

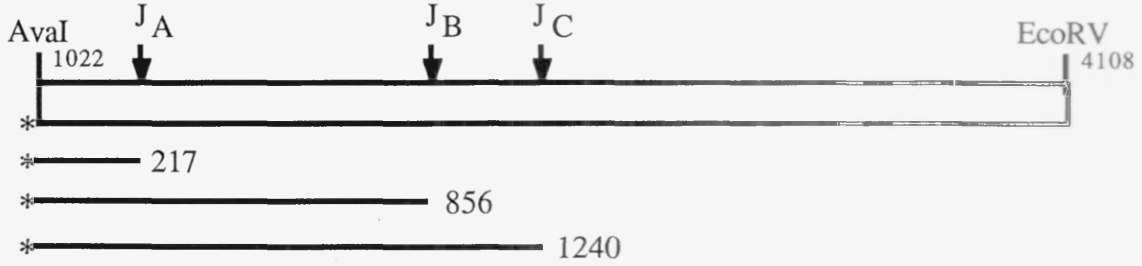
Figure 5.1.

Nuclear and cytoplasmic distribution of Py RNAs in 82-Rat. A 3' end-labelled probe was constructed (lower right) from an *Ava*I-*Eco*RV fragment of Py DNA (white box) extending between nts 1022 and 4108. Early regions A, B and C of 82-Rat (upper right) are contiguous with the probe to different extents 3' to the labelled *Ava*I site. The points at which transcripts from the three early regions first lose complementarity to the probe are indicated (J_A, J_B and J_C respectively). Thus transcripts from early region A protect a probe fragment 217 nts in length against S1 nuclease digestion (lower right), while transcripts from early regions B and C give protected fragments of 856 and 1240 nts respectively. The left-hand panel shows the S1-resistant products obtained after annealing the probe with:

Lane 1: 82-Rat total nuclear RNA. Lanes 2 and 3: 82-Rat total cytoplasmic RNA (20 and 50 μg respectively). Lane 4: yeast RNA (50μg). Lane 5 contains size markers (end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in pAT153) and lane 6 contains an aliquot of the undigested probe. The amount of nuclear RNA used in lane 1 was not quantified spectrophotometrically, because of the likelyhood of DNA contamination, but was instead calculated to be equivalent in terms of cell number to the 50μg of cytoplasmic RNA in lane 3. S1-resistant products were fractionated by alkaline agarose gel electrophoresis.


1 2 3 4 5 6
Probe
1240
856
217
1452
1295
656
559
464


82-Rat
early A
late A
late B
early B
late C
early C
early B'
J A
J B
J C


AvaI
1022
J A
J B
J C
EcoRV
4108
217
856
1240

RNAs, which protect an 856 nt. fragment in the 3' S1 analysis (Fig. 5.1), are present in the cytoplasmic fraction; this result gives a measure of the efficiency with which transcripts polyadenylated at the Py major early site are transported into the cytoplasm and/or a measure of the stability of such cytoplasmic mRNAs. The proportion of the early region A transcripts present in the cytoplasm is even greater; the abundance of these RNAs is indicated by the intensity of the 217 nt. band (Fig. 5.1). In contrast, the majority of the early C transcripts, which give a 1240 nt. protected fragment, are present in the nuclear fraction. Comparison of the abundances of nuclear species shows that, while early B and early C transcripts are present at near-identical levels, the relative abundance of early A nuclear transcripts is reduced. Conversely, early A transcripts are the most abundant cytoplasmic Py RNAs; while the level of early B cytoplasmic RNAs is not much lower, there is very little cytoplasmic RNA derived from early region C. Cleavage at the alternative poly(A) site of transcripts derived from any of the three partial early regions would result in protected fragments of ~475 nts. The signal in this region corresponds to only a very low level of alternative poly(A) site use in 82-Rat cells (Fig. 5.1, track 3).

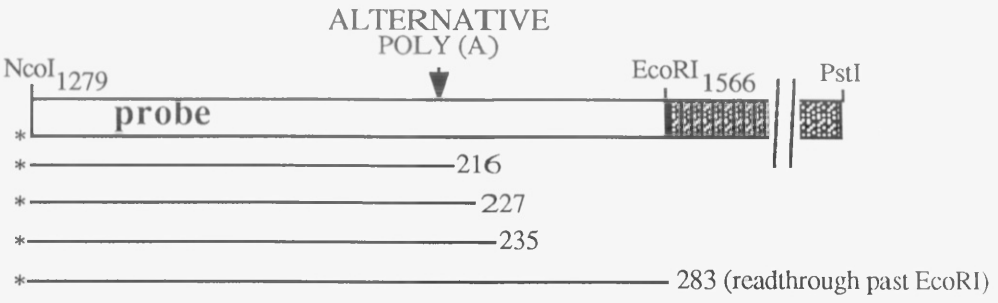
Polyadenylation and nuclear-cytoplasmic distribution of Py RNAs in the tsC3T3 lines.

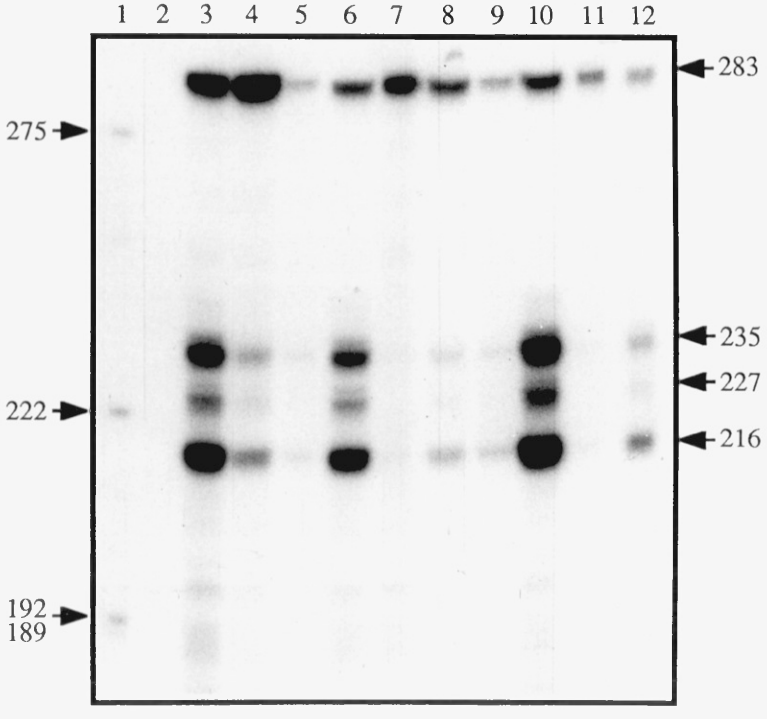
3' S1 nuclease analysis was performed in order to establish the extent to which the Py alternative poly(A) site is used in the parental transformed line tsC3T3U and in the single-cell clones tsC3T3-1, -2 and -3 (Fig. 5.2). For tsC3T3U, -2 and -3 the majority of cytoplasmic RNAs initiated at the Py early promoter are polyadenylated at the alternative site (Fig. 5.2, lanes 6, 10 and 12); the pattern of processing closely resembles that of 53-Rat (Fig. 5.2, lane 3). The relatively low level of total signal in the tsC3T3-3 lane (Fig. 5.2, lane 12) may reflect a correspondingly low level of Py mRNAs in this line, although it is possible that this RNA sample was partially degraded. The Py RNA processing pattern in tsC3T3-1 cells is somewhat different. Comparison of 53-Rat and tsC3T3-1 total

Figure 5.2.

Nuclear-cytoplasmic distribution of Py RNAs in tsC3T3 variant cell lines. The 3' end-labelled probe for this experiment is represented at the top of the figure. Py sequences are shown in white, and plasmid sequences are shaded. The probe extends from the labelled *Nco*I site at Py nt. 1279 through to the *Eco*RI site at nt. 1566; the remainder of the probe consists of plasmid vector (pxf3) sequences, and is consequently not complementary to Py RNAs. RNAs cleaved at the alternative poly(A) site are predicted to give protected fragments in the range 216-235 nts as shown; transcripts extending beyond the *Eco*RI site protect a single fragment of 283 nts The lower panel shows S1-resistant products obtained after annealing with:

Lane 2: Rat-1 total cell RNA. Lane 3: 53-Rat total cell RNA. Lane 4: tsC3T3-1 total cell RNA. Lanes 5 and 6: nuclear and cytoplasmic RNAs from the tsC3T3 parental line, tsC3T3U. Lanes 7 and 8: nuclear and cytoplasmic RNAs from tsC3T3-1. Lanes 9 and 10: nuclear and cytoplasmic RNAs from tsC3T3-2. Lanes 11 and12: nuclear and cytoplasmic RNAs from tsC3T3-3. 40μg RNA was used for lanes 2, 3, 4, 6, 8 and 12; RNA from equivalent cell numbers was used for lanes 5, 7, 9 and 11. Lane 1 contains markers (end-labelled *Dde*I fragments of a *Bam*HI-linear clone of Py in pAT153) whose sizes are indicated to the left.


ALTERNATIVE
POLY (A)
NcoI 1279
EcoRI 1566
PstI
probe
216
227
235
283 (readthrough past EcoRI)


1 2 3 4 5 6 7 8 9 10 11 12
283
275
235
227
222
216
192
189

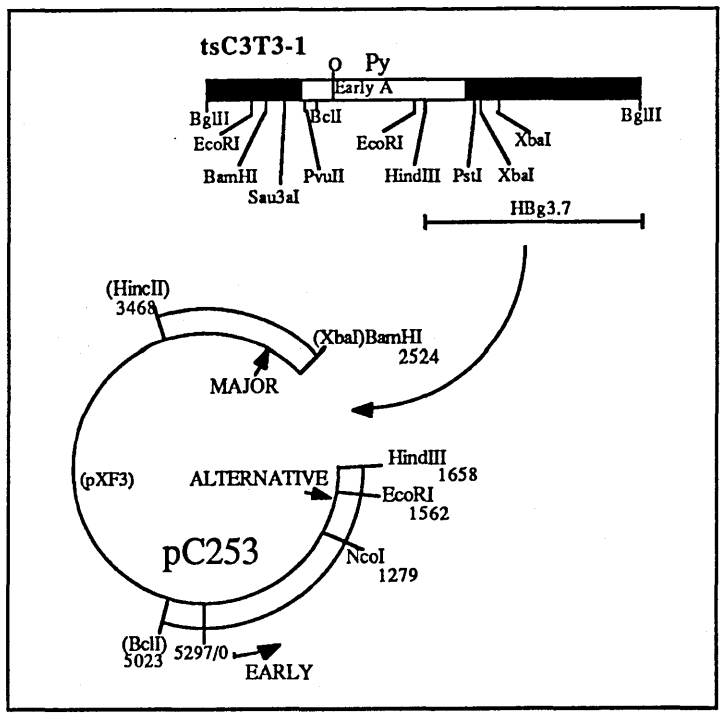
cell RNAs (Fig. 5.2, lanes 3 and 4) shows that the latter line has a lower level of alternative poly(A) site use, while a relatively large amount of Py RNA extends beyond the viral *Eco*RI site. As judged from examination of tsC3T3-1 nuclear and cytoplasmic RNAs (Fig. 5.2, lanes 7 and 8), the majority of these RNAs extending beyond the *Eco*RI site are found in the nuclear fraction (as evidenced by the abundant 238 nt. protected fragment in lane 7 of Fig. 5.2).

Plasmid reconstruction of the 82-Rat early region C and tsC3T3-1 transcription units.

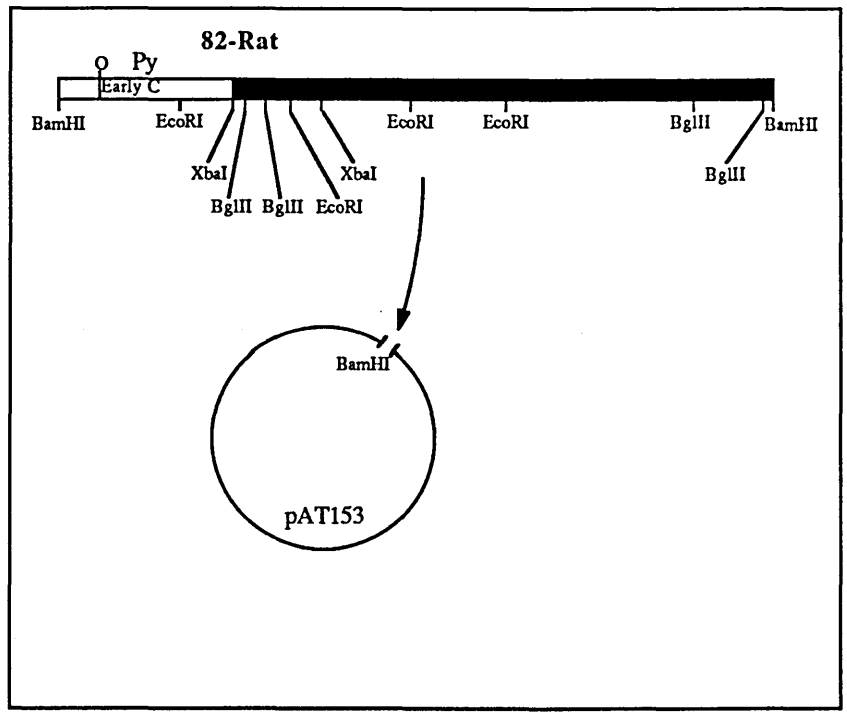
If the characteristic transcriptional patterns of tsC3T3-1 and the 82-Rat early region C reflect the nature of sequences downstream from each viral insert, then these patterns might be reproduced in plasmid reconstructions containing the cloned 3' flanking sequences. This approach might then be used to idntify sequence elements responsible for the transcriptional phenotypes. Plasmid derivatives of tsC3T3-1 and the 82-Rat early region C are shown in Fig. 5.3. For tsC3T3-1 the HBg3.7 fragment was inserted into *Hin*dIII- and *Bam*HI-cleaved pC253, giving pCtsC-1; the arrangement of viral and cellular sequences as far as the *Bgl*II site in pCtsC-1 is thus exactly the same as that in the parental cell line, with in addition the sequences including the Py major early poly(A) site downstream (Fig. 5.3). If sequences within the HBg3.7 fragment destabilise, or prevent the formation of, transcripts reading through to the major poly(A) site, relatively little RNA extending beyond the Py *Eco*RI site will be detected in the cytoplasm of transfected cells. The latter RNA species would protect a ~283 nt. fragment of the 3' end-labelled probe used (Fig. 5.4). In cases where polyadenylation at the (upstream) alternative Py site is favoured, increased intensity of a triplet of protected fragments in the range ~ 216 to ~235 nts would be detected. For 82-Rat, the entire λ82-R insert was cloned into the *Bam*HI site of pAT153, giving p82R; this derivative was used for transient expression. In this case no known poly(A) site was placed downstream from the cellular sequences;

Figure 5.3.

Plasmid derivatives of tsC3T3-1 and 82-Rat early region C. The left-hand part of the figure shows the construction of derivative pCtsC-1. The HBg3.7 fragment from λtsCBc-2 (Fig. 4.1) was inserted into *Hin*dII- and *Bam*HI-cleaved pC253 (Fig. 3.7) as shown. The resulting plasmid (pCtsC-1) contains Py (white box) early region sequences essentially identical to those of tsC3T3-1, together with the viral-host junction and some 3 kb of flanking mouse DNA (black box). The right-hand part of the figure shows the sequences of the λ82-R insert (Fig. 4.1), which were inserted into the *Bam*HI cloning site of pAT153, forming p82R.



Construction of pCtsC-1

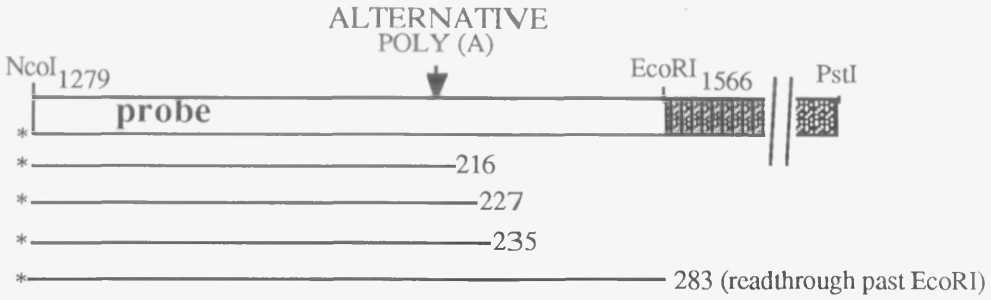


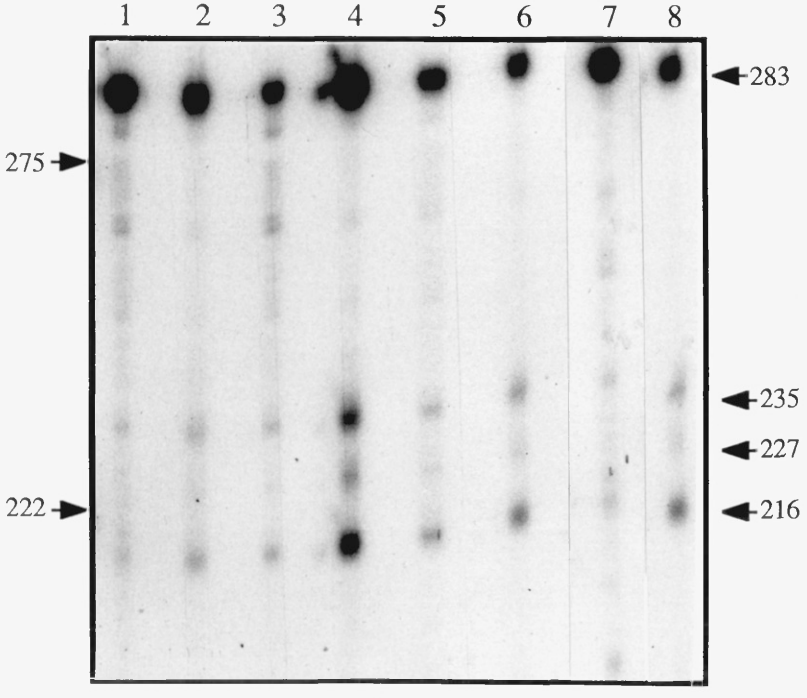
Construction of p82-R

Figure 5.4.

Nuclear-cytoplasmic distribution of RNAs transiently expressed from pCtsC-1 and p82R. The 3' end-labelled probe for this experiment is represented at the top of the figure. Py sequences are shown in white, and plasmid sequences are shaded. The probe extends from the labelled *Nco*I site at Py nt. 1279 through to the *Eco*RI site at nt. 1566; the remainder of the probe consists of plasmid vector (pxf3) sequences, and is consequently not complementary to Py RNAs. RNAs cleaved at the alternative poly(A) site are predicted to give protected fragments in the range 216-235 nts as shown; transcripts extending beyond the *Eco*RI site protect a single fragment of 283 nts The lower panel shows S1-resistant products obtained after annealing with:

Lanes 1 and 2: nuclear and cytoplasmic RNAs transiently expressed from p82R. Lanes 3 and 4: nuclear and cytoplasmic RNAs transiently expressed from pCtsC-1. Lanes 5 and 6: nuclear and cytoplasmic RNAs transiently expressed from pCFR1 (see Fig. 3.8). Lane 7: total cell RNA transiently expressed from the 'empty' vector pC253 (Fig. 3.7). Lane 8: total cell RNA from 53-Rat. 40μg RNA was used for lanes 2, 4, 6, 7 and 8; RNA from equivalent cell numbers was used for lanes 1, 3 and 5. The positions to which markers of the sizes indicated migrated are shown to the left.


ALTERNATIVE
POLY (A)
NcoI 1279
probe
EcoRI 1566
PstI
216
227
235
283 (readthrough past EcoRI)


1
2
3
4
5
6
7
8
283
275
235
227
222
216

however, if sequences contained within the λ82-R clone are sufficient to elicit the 82-Rat early region C phenotype, then very little of the RNA transiently expressed from the p82R plasmid should be found in the cytoplasmic compartment.

Results obtained using the 82-Rat and tsC3T3-1 derivatives are equivocal. A considerable proportion of the RNA transiently expressed from p82R is cytoplasmic. There may, however, be a slight increase in the ratio of nuclear:cytoplasmic transcripts for p82R (compare the intensity of the ~283 nt. fragment in lanes 1:2 and 3:4 of Fig. 5.4).

For the tsC3T3-1 derivative, pCtsC-1, the majority of cytoplasmic transcripts extend beyond the Py *Eco*RI site, as they do in the parental cell line (compare Fig. 5.4, lane 4 with Fig. 5.2, lane 8). Unlike the parental cell line, however, the tsC3T3-1 plasmid derivative does not seem to give rise to elevated levels of nuclear transcripts (compare Fig. 5.4, lane 3 with Fig. 5.2, lane 7).

DISCUSSION

There is considerable variation in the pattern of processing of viral RNAs among the Py-transformed cell lines which have been examined. For the 53-Rat line, the distinctive pattern of RNA processing has been shown to result, in part at least, from the nature of cellular sequences adjacent to the site of viral integration (see chapter 3). This investigation was extended to the transformed line 82-Rat, and to a series of clones, tsC3T3-1, -2 and -3, derived from a single Py-transformed line.

82-Rat presents a complex transcriptional pattern, having three separate actively transcribed early regions (Figs 4.2, 5.1). Early regions A and C are truncated by recombination with cellular sequences at either end of the viral insert. The abundances of cytoplasmic RNAs derived from these truncated early regions were respectively slightly higher and much lower than the abundance of cytoplasmic RNAs derived from early

region B, which is contained within the body of the viral insert (Fig. 5.1). While the levels of steady-state nuclear RNA derived from early regions B and C were similar, less nuclear RNA corresponding to early region A was observed. If one assumes a roughly equal level of transcription initiation for the three 82-Rat early regions, then one possible interpretation of these data is that early region A mRNAs are transported more efficiently to the cytoplasm than are early region B mRNAs. Thus the cellular signals used for early region A polyadenylation (B. Davies, personal communication) and/or the other cellular sequences present in early region A transcripts are more efficient in directing export from the nucleus than are the wild-type Py sequences used by early region B. This result further demonstrates the value of the truncated Py early regions present in transformed cells in the identification of cellular gene expression sequences. The very low level of cytoplasmic early region C transcripts could result from instability of such species in the cytoplasm, although early region C transcripts would appear to be as stable in the nucleus as those derived from early region B. It is perhaps more likely, then, that early region C transcripts, unlike those of early region A, are not efficiently polyadenylated at a site in the 3' flanking sequences (certainly, using S1 analyses no discrete nuclear or cytoplasmic RNA 3' termini were detected in the region of cellular sequences up to ~9 kb downstream from early region C). If this is the case then an intriguing distinction can be drawn between the 82-Rat early region C and the early regions present in 53-Rat and tsC3T3-2, whose viral-host fusion transcripts also fail to use a cellular poly(A) site, but instead use the Py alternative site. No appreciable level of alternative site use was observed for any of the 82-Rat partial early regions. Transformed lines containing actively transcribed Py early regions which, in the absence of an efficient viral or cellular site, do not use the alternative poly(A) site would add weight to the argument that the alternative site is specifically activated in lines such as 53-Rat and tsC3T3-2.

Attempts to reproduce in a plasmid-based assay the characteristic transcription pattern of the 82-Rat early region C were not wholly successful; considerable amounts of

cytoplasmic RNA were generated from a plasmid containing early region C with approximately 9 kb of the downstream cellular sequences. The proportion of total Py RNA present in the cytoplasmic fraction was, however, slightly lower for the 82-Rat plasmid than for equivalent plasmids containing flanking sequences from tsC3T3-1 or 53-Rat. If the nuclear-cytoplasmic distribution of early region C RNAs is determined solely by particular 3' flanking cellular sequences, it would seem that these sequences do not lie within the p82R plasmid (Fig. 5.3). This plasmid was as efficient as a plasmid containing the entire viral early region when tested for focus forming ability on Rat-1 cells; this observation further suggests that p82R transcripts are transported to the cytoplasm more efficiently than those of the 82-Rat early region C. Sequences more than 9 kb downstream from the 82-Rat Py insert might be acting in a negative manner to prevent the formation of stable cytoplasmic early region C transcripts. An alternative explanation is that the early region C transcription pattern results from some as yet undefined characteristic of the 82-Rat cell line, or some feature of the chromatin of which early region C is a part. In either case the phenomenon might not be reproducible following transfection of plasmid derivatives into Rat-1 cells.

The structure of the Py insert in tsC3T3-1 cells has been shown to differ considerably from those of other lines derived from the same transformed focus (see chapter 4). This difference in genomic organisation was found to be reflected in a difference of transcriptional pattern when RNA from tsC3T3-1 was compared with RNAs from the other, apparently related tsC3T3 lines (Fig. 5.2). Processing of tsC3T3-1 RNAs would appear to share some of the features of both the tsC3T3-2 (or 53-Rat) and the 82-Rat early region C transcription patterns. Thus, while a considerable proportion of tsC3T3-1 transcripts are found in the nucleus, the alternative poly(A) site is also used to a certain extent, though not to the extent to which this site is used in tsC3T3-2 and 53-Rat cells (Fig. 5.2). Plasmid reconstructions of the type used in the analysis of the 53-Rat transcription pattern (chapter 3) may be of use in the further characterisation of cellular

sequences responsible for the diverse Py transcription patterns described here.

CHAPTER 6: MATERIALS AND METHODS

MATERIALS

REAGENTS AND EQUIPMENT.

Acrylamide and NN'-methylene-bisacrylamide	BDH. Stock solutions were deionised with ion exchange resin (Bio-Rad AG501-X8) and filtered before use.
Agarose	SeaKem ME.
Ampicillin	Penbritin (Beecham Research Laboratories).
Bacteriophage λ packaging mixes	Stratagene Cloning Systems.
Caesium chloride	Koch-Light.
Chromatography	G75 Sephadex: Pharmacia. Oligo(dT)-cellulose: Collaborative Research.
Detergents	Sarkosyl (N-dodecyl sarcosinate, 30% solution) and SDS (sodium dodecyl sulphate): BDH.
Dialysis Tubing	Spectrapor.

EN^{3}HANCE	NEN.
Ethylacetate	Fisons AR quality .
Ficoll	Pharmacia (400k).
Films	Polaroid type 667 for photography of EtBr-stained agarose gels, Kodak XAR-5 and Fuji RX for autoradiography.
Formaldehyde	Fisons (38% w/v solution).
Formamide	BDH (specially pure, no. 44254) was used for RNA solution hybridisation and Fisons (A.R. grade) was used for filter hybridisation. Both were deionised using an ion exchange resin (Bio-Rad AG501-X8) and filtered before use.
Guanidinium isothiocyanate	Fluka (purum grade).
Nitrocellulose	BA85, pore size: 0.45 μm : Schleicher and Schuell.
Nucleic acids	M13 mp18 and mp19 rf. DNA: P-L. M13 primer (15mer): New England Biolabs. λgt10: Vector Cloning Systems. Restriction enzyme linkers: Collaborative Research.

	Dideoxyribo- and deoxyribo-triphosphates: P-L.
Radiochemicals	Isotopically labelled chemicals were obtained from Amersham International or NEN. Their specific activities are referred to in the text.
RNAsin	Human placental ribonuclease inhibitor: Bolton Biologicals.
Vanadyl Ribonucleoside Complexes	BRL (Speciality Reagent).
X-Ray intensifying screens	Du Pont Lightning Plus.

All other reagents and equipment were obtained from Sigma or ICRF Central Services.

ENZYMES.

Avian myeloblastosis virus RNA directed DNA polymerase (reverse transcriptase)	Anglian Biotechnology.
Calf intestinal alkaline phosphatase	Boehringer.
DNA restriction enzymes	New England Biolabs, BRL, Boehringer and NBL.

DNase I	Worthington. This was further purified free of RNase activity by affinity chromatography as described in Maniatis *et al.* (1982).
DNA polymerase I	Boehringer.
E.coli DNA ligase	New England Biolabs.
Klenow DNA polymerase	B R L .
Lysozyme	Sigma.
Nuclease S1	Sigma.
Proteinase K	Sigma.
RNase A	Sigma.
RNase H	BRL.
T4 DNA ligase	New England Biolabs.
T4 DNA polymerase	P-L.
T4 polynucleotide kinase	Pharmacia.

CELLS AND VIRUS STOCKS

Bacteria	Strains specified in the text are described in detail elsewhere (Maniatis *et al.* 1982).
Rat-1	Immortal rat fibroblast (Freeman *et al.* 1973).
Mouse 3T6	Swiss albino mouse cells (ATCC).
53-Rat, 82-Rat, 7axB, 7axT	Rat cell lines obtained by infection of Rat-1 cells by wild-type (strain A2) polyomavirus (Lania *et al.* 1979).
tsC-3T3-1, 2, 3, wt3T3-1	Single cell clones derived from a parental line obtained by infection of Balb/c-3T3 cells by Py (tsC or wild-type (wt) A2) and passage at the nonpermissive temperature in the presence of anti-viral serum (M. Read, personal communication, Kamen *et al.* 1979).

All cell lines and viruses (Py A2 and tsA strains) were obtained from laboratory stocks kindly provided by M. Read.

FORMULATION OF FREQUENTLY USED SOLUTIONS

PBS.A	1% w/v NaCl, 0.025% w/v KCl, 0.14% w/v Na_2HPO_4, 0.025% w/v KH_2PO_4.

PBS	10mM $CaCl_2$, 6mM $MgCl_2$ in PBS.A.
Versene	0.54mM EDTA in PBS.A.
TE	1mM EDTA, 10mM Tris.Cl pH7.4.
SM	10mM Tris.Cl pH7.4, 10mM $MgCl_2$, 100mM NaCl 0.05% w/v gelatin.
1xTBE	89mM Tris-base, 89mM Boric acid, 2mM EDTA.
1xMOPS-E	20mM MOPS, 1mM EDTA, 5mM NaOAc pH adjusted to 7.0 with glacial acetic acid.
GS1	2M NaCl, 0.5M NaOH.
GS2	2M NaCl, 0.5M Tris-base, pH adjusted to 8.0 with HCl.
1xSSC	150mM NaCl, 15mM sodium citrate, pH adjusted to 7.5 with NaOH.
1xDenhardts solution	0.02% w/v BSA, 0.02% w/v Ficoll 400, 0.02% w/v polyvinylpyrrolidone.
2FC	49.5% v/v redistilled phenol, 49.5% v/v chloroform, 1% v/v isoamyl alcohol, 0.1% w/v 8-hydroxyquinoline saturated with TE pH 8.0.

STET	50mM Tris.Cl pH 8.0, 50mM EDTA pH 8.0, 8% w/v sucrose, 5% v/v Triton X-100.
Sucrose gradient solutions	40% or 10% w/v sucrose, 5mM EDTA, 20mM Tris.Cl pH8, 1M NaCl. For plasmid gradients 25μg/ml ethidium bromide was also present, but this was omitted from sucrose solutions used for fractionating genomic DNA.
Formamide stop buffer	10mM NaOH, 1mM EDTA, 80% v/v formamide 0.02% w/v bromophenol blue, 0.02% w/v xylene cyanol.

FORMULATION OF FREQUENTLY USED BACTERIAL GROWTH MEDIA

L Broth	1% w/v Bacto-tryptone, 0.5% w/v Bacto-yeast extract, 1% w/v NaCl.
2xTY	1% w/v Bacto-tryptone, 1% w/v Bacto-yeast extract 0.5% w/v NaCl.

For growth on solid media, L broth or 2xTY were supplemented with 1.5% w/v Bacto-agar. For bacteriophage λ plaque-lift assays L broth supplemented with 0.7% agarose was used to prevent the adhesion of the overlay to nitrocellulose. For M13 overlays the following was used:

H top	1% w/v Bacto-tryptone, 0.8% w/v NaCl, 0.8% w/v Bacto-agar.

METHODS

Mammalian Cell Culture

All mammalian cell lines were grown in a 10% CO_2: 90% air humidified incubator, using Dulbecco-modified Eagles medium (DMEM) supplemented with 0.1% w/v glutamine and 5% v/v FCS. Cells were grown on plastic tissue culture dishes (9cm, circular) or 24.5 x 24.5 cm bioassay trays (Nunc). Cells were removed from the dishes, following an initial versene wash, with a 3:1 mixture of versene:0.1% w/v trypsin in Tris saline, and passaged at appropriate densities.

Viral Infection of 3T6 Cells.

Mouse 3T6 fibroblasts ($5x10^7$/bioassay tray) were infected at a multiplicity of 50 pfu/cell in 10 ml serum-free DMEM with frequent rocking in a 37°C incubator. The incubator was flushed periodically with CO_2 in order to maintain a relatively acid pH. After 90 mins, the cells were refed with 100 ml DMEM + 5% horse serum and incubation was continued at 37°C.

Preparation of High Molecular Weight Genomic DNA.

Eukaryotic cell DNA was prepared by a variation of the procedure of Blattner *et al.* (1978). Cells were cultivated on large bioassay trays until almost confluent. After removal of the medium, cells were washed twice in PBS and drained thoroughly. Each tray of cells was then overlaid with 4ml of lysis buffer (10mM TrisHCl pH8, 10mM EDTA pH8, 5mM NaCl, 0.5% w/v SDS and 0.1mg/ml proteinase K) and incubated at 37°C for 10 mins. The lysed mixture was then carefully placed in a 50ml Falcon tube, using a rubber policeman, and incubated at 37°C for a further 3-12 hr. The lysed cells were then extracted twice with 2FC and once with chloroform. The DNA was subsequently dialysed against 2l 1M NaCl, TE for 12 hrs at 4°C and then 4l TE for 48 hrs at 4°C. The nucleic

acid was then treated with 20μg/ml RNase A at 37°C for 3 hrs, protein extracted as above and dialysed against 4l TE for 48 hrs at 4°C before storage at 4°C.

Isolation of Mammalian Total Cellular RNA and Poly $(A)^+$ Selection.

Prior to lysis cells were washed twice with PBS and thoroughly drained. The cells were lysed at RT with 4M GT (50% w/v guanidinium isothiocyanate, 0.5% v/v Sarkosyl, 25mM HEPES, 100mM β-mercaptoethanol, pH adjusted to 7 with NaOH and filtered) using 5ml per bioassay dish (Chirgwin *et al.*, 1979). High molecular weight DNA was sheared by vortexing, and passage of the extract through a 21 gauge needle. The extract was then layered onto an equal volume cushion of 5.7M CsCl, 25 mM NaOAc, and centrifuged at 200,000g (35,000 rpm)/20°C/20-24 hrs in a Beckman SW40 rotor. After centrifugation the supernatant was discarded and the tubes were sliced through just above the RNA pellet to prevent contamination from adherent protein or DNA. The pellets were transferred to new tubes and washed with 70% v/v ethanol. RNA was redissolved in RB (10mM Tris.Cl pH7.4, 5mM EDTA, 1% w/v SDS) and extracted with an equal volume of 80% v/v chloroform, 20% v/v butan-1-ol. The organic phase was re-extracted with RB, the two aqueous phases combined and NaOAc pH5.4 added to 0.3M. RNA was then ethanol precipitated, resuspended in 0.3M NaOAc pH5.4, transferred to an Eppendorf tube and reprecipitated. The RNA was either stored in water, or under ethanol, at -20°C. Where RNA had been prepared from cells which had been transfected with plasmid sequences it was treated with RNase free DNase I before further use (see below). RNA was fractionated into $poly(A)^+$ and $poly(A)^-$ components by chromatography on oligo(dT)-cellulose as described by Maniatis *et al.*(1982). The yield of $poly\ (A)^+$ RNA varied from 2-5% of the total RNA. When the amount of RNA present was low, ethanol precipitation of the RNA from 0.3M NaOAc pH 5.4 was assisted by the addition of 1-2μg of RNase free dextran.

Preparation of Nuclear and Cytoplasmic RNA.

Cells were washed twice with ice-cold PBS and scraped into a corex tube using a rubber policeman bung. After pelleting (2,000g, 5min., 4°C), excess PBS was removed and cells were resuspended in ice-cold lysis buffer (0.14M NaCl, 1.5mM $MgCl_2$, 10mM Tris-HCl pH 8.6, 0.5% NP40, 10mM VRC: 6ml per bioassay tray). After lysis (5 min. on ice), the suspension was loaded onto an equal volume cushion of 24% sucrose (w/v) in lysis buffer in a sterile Beckman 331374 tube. After centrifugation at 50,000g (10,000 rpm)/4°C/20 mins/SW40, the upper (cytoplasmic) fraction was carefully removed and added to an equal volume of complete 2x P.K. buffer (300mM NaCl, 25mM EDTA, 2% w/v SDS, 200mM Tris-HCl pH7.5, 200µg/ml Proteinase K). After careful removal of the sucrose, the nuclei were resuspended well in lysis buffer at 0°C (6ml per bioassay tray) and an equal volume of complete 2x P.K. buffer was slowly added. Both fractions were incubated for 30 mins at 37°C: the viscosity of the nuclear solution was reduced by repeated shearing through a 19 gauge syringe needle. After extraction with 2FC nucleic acids were recovered by ethanol precipitation. DNA was removed from the nuclear fraction by incubation with pancreatic DNAse I (2µg/ml) in 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 2mM VRC for 30 mins at 37°C. The solution was brought to 10mM EDTA, 0.1% SDS, phenol extracted and the RNA recovered by ethanol precipitation (this step was included in the preparation of cytoplasmic RNA if no fractionation on oligo(dT)-cellulose was performed).

DNA Quantitation and Manipulation.

DNA and RNA were quantitated by determining the absorbance of dilutions of stock solutions at 260nm (1 OD_{260} = 50µg/ml DNA, 40µg/ml RNA). Following enzyme treatment, plasmid and cellular DNA was normally phenol extracted and ethanol precipitated before further manipulation (although restriction digests of plasmid DNA for electrophoresis would normally be loaded directly onto gels following the addition of

dye). If small quantities of DNA were present 1-2μg of yeast RNA, or dextran where the absence of RNA was important, was added to assist ethanol precipitation. In addition, in the absence of any salt, either NaCl to 1M, or NaOAc pH5.4 to 0.3M, was added to facilitate DNA precipitation. Restriction enzyme digestions were performed as described in Maniatis *et al.* (1982), except that 5 units of enzyme/μg of DNA was used. To digest cellular DNA the buffers and enzymes were initially mixed in gently to prevent shearing. After 3 hr at 37°C, more restriction enzyme was added to the cellular DNA and the incubation continued for a further 2 hrs. Intermolecular ligations were performed in small volumes, generally 10μl for a total DNA content of 1μg. For ligations involving plasmid sequences a 1:10 ratio of plasmid:insert DNA was used. When the plasmid was capable of self-ligation, the compatible ends were dephosphorylated before use (see below). Ligations involving λ vectors were normally set up at three different ratios of arms:insert (1:2, 1:1 and 2:1 for the ratio of compatible ends arms:compatible ends insert) and after packaging and titration the most suitable ratio was selected for construction of a particular library. Ligations were performed overnight at 4-15°C in ligation buffer (25mM Tris.Cl pH7.4, 10mM $MgCl_2$, 1mM spermidine, 1mM ATP, 10mM DTT, 0.1mg/ml BSA) and proteins were not extracted before further use. Dephosphorylation of DNA was performed using calf intestinal alkaline phosphatase as described in Maniatis *et al.* (1982). Following incubation at 37°C the reaction was heated at 65°C for 10 min, and then extracted twice with 2FC.

Bacteriophage λ Vectors.

The λ replacement vectors EMBL4 (Frischauf *et al.*, 1983) and L47.1 (Loenen and Brammar, 1980) were propagated on the *E. coli* strains WL87 or LE392 (recombinant and wild type lysis) and WL95 or P2392 (recombinant lysis only). The λ insertion vector λgt10 (Huynh *et al.*, 1985) was used for cloning cDNA sequences. The wild type vectors were propagated on *E. coli* C600 (recombinant and wild type lysis), whilst C600

HFL (recombinant lysis only) were used not only to assess the number of recombinants present in the cDNA libraries, but also for the screening and propagation of recombinants. Bacteria harbouring λ phage were grown in L broth supplemented with 20mM $MgSO_4$, and plated in top agarose onto L broth agar plates. Phage stocks were derived from the supernatants of lysed bacteria. Alternatively a lawn of bacteria, which had undergone almost complete lysis, was overlaid with SM and rocked at RT for 1-2 hrs before the SM was removed for storage. Both plate lysates and liquid lysates were stored over chloroform at 4°C.

Size Fractionation of Restricted Total Cellular DNA.

In order to enrich for a genomic restriction fragment harbouring integrated Py sequences, digests of total cell DNA (500μg) were loaded onto 10-40% sucrose gradients and spun at 120,000g (24,000 rpm)/10°C/20 hrs/SW27. 1ml fractions were collected via a 19 gauge needle inserted through the base of the tube, and the approximate position of the size fraction of interest was determined after running 50μl aliquots of every tenth fraction on a 0.8% mini agarose gel. 400μl aliquots of each fraction in the appropriate range were precipitated, run out on 0.8% agarose gels and blotted onto nitrocellulose. The fraction(s) containing Py-specific sequences were then identified by hybridisation to Py probes as described below.

Bacteriophage λ Libraries: Packaging and Screening.

Ligations performed with EMBL4, L47.1 or λgt10 were packaged using the protocol recommended by the manufacturers of the packaging mixes. Libraries were plated out on large bioassay trays to obtain about 50,000 plaques per plate, although a maximum of 40% of any particular packaging reaction was used per plate. A 50ml overnight culture of the appropriate bacteria was spun down and resuspended in 20ml of ice cold 10mM $MgSO_4$. The required fraction of the packaging mix was added to 1ml of SM plus 1ml of

the bacteria in a plastic universal and incubated at 37°C for 15 mins without shaking. The universal was then filled with molten top agarose at 42°C, mixed and quickly poured onto a bioassay tray containing 400ml of pre-warmed L broth agar at 37°C. Once set the plates were inverted and incubated for 12 hrs at 37°C. Plates were then transferred to 4°C for 2 hrs prior to the taking of nitrocellulose lifts, to allow the top agarose to harden. Phage DNA lifts were taken in duplicate, to guard against false positives, and the filters and plates marked with ink to facilitate later alignment. Nitrocellulose was left on the plates for 30 secs for the first lift, and for 60 secs for subsequent lifts. Subsequently the sheets were placed in a bath of GS1 for 30 secs, GS2 for 2-5 mins and finally rinsed in 2xSSC before being air dried and baked under vacuum at 80°C for 2 hrs. Conditions for hybridisation and washing of the filters are presented below. Positive plaques were picked and purified to homogeneity before further use.

Small Scale Preparation of Bacteriophage λ DNA.

A single well isolated plaque was picked directly into 10ml of a 1 in 100 dilution of a saturated culture of the appropriate host bacteria. Lysis of the culture was generally complete after 12 hrs shaking at 37°C. The purification of phage DNA from the liquid lysate was as described by Maniatis *et al.* (1982). Enough DNA for 1-5 restriction digestions, carried out in the presence of 50μg/ml RNase A, was normally obtained.

Large Scale Preparation of λ DNA.

A 10ml overnight culture of the appropriate bacteria was spun down and resuspended in ice cold 4ml 10mM $MgSO_4$. These were then incubated with approximately $5x10^7$ phage for 15 min at 37°C and then used to inoculate 400ml of prewarmed L broth plus 20mM $MgSO_4$ which was then shaken vigorously overnight. If lysis had failed to occur after 12 hrs, the bacteria were diluted 1 in 4 and shaking continued.The purification of phage DNA from the liquid lysate by centrifugation through CsCl step gradients was as described by Maniatis *et al.*, (1982).

Plasmid Vectors.

Plasmid vectors were either pAT153 (Twigg and Sherratt, 1980) or pxf3 (Hanahan, 1983). Recombinant plasmids were propagated on *E. coli* strains DH1 or DH5 (see Maniatis *et al.*, 1982), which were grown in L broth supplemented with 50 μg/ml ampicillin. In the absence of plasmids these bacteria were grown in L broth alone. For long term storage of bacteria, fresh overnight cultures were stored at -20°C after the addition of glycerol to 20% v/v.

Preparation, Storage and Transformation of Competent Bacteria.

Competent bacteria were prepared, stored and transformed exactly as described (Scott and Simanis, 1985). The identification of bacteria containing recombinant plasmids was usually performed by the restriction enzyme analysis of plasmid DNA. However, when a high background of non-recombinant pAT153 or pxf3 was present, it was sometimes possible to identify recombinants by replica plating onto L broth agar plates containing both ampicillin and 12.5 μg/ml tetracycline.

Small Scale Preparation of Plasmid DNA.

The method used was a modification of that of Holmes and Quigley (1981). Briefly, 1.5 ml of a saturated (overnight) culture derived from a single transformant colony was spun for 2 mins (12,000g, 4°C). After removal of the supernatant by aspiration the cell pellet was resuspended in 100μl STET + 2μg/ml lysozyme (fresh). Chromosomal DNA was denatured by boiling for 45 secs., and was pelleted by centrifugation (15 mins, 12,000g, 4°C). The supernatant was transferred to a fresh tube containing 100μl isopropanol. After chilling at -20°C for 10 mins, nucleic acids were pellleted by centrifugation (5 mins, 12,000g, 4°C) and washed with 70% ethanol. Dried pellets were resuspended by incubation (5 mins, 37°C) with 25μl TE+ 0.1 mg/ml RNAse A: 5μl aliquots were used directly in restriction enzyme analysis, while the remainder was stored

at -20 °C.

Large Scale Preparation of Plasmid DNA (Holmes and Quigley, 1981).

An 10 ml overnight bacterial culture was used to inoculate 400 ml of prewarmed medium in a 2l flask. At OD_{600} of 0.4, spectinomycin was added to 300 μg/ml and the culture continued overnight. Cells were pelleted by centrifugation at 10,000g for 10 mins, and washed twice in 15 ml of 150 mM NaCl, TE (by suspension and repelleting) during which they were transferred to a 30 ml Corex tube. The final pellet was resuspended, by vortexing and pipetting, in 10 ml STET containing 2mg/ml lysozyme (freshly added), and immediately placed in a boiling water bath for 2 mins. The mixture was then cooled on ice prior to centrifugation at 200,000g (30,000 rpm)/10°C/30 mins in a Beckman SW40 rotor. Plasmid DNA was purified from the resulting supernatant by caesium chloride density centrifugation (as described by Maniatis *et al.*, 1982) in either a Beckman 50Ti rotor at 140,000g (47,000 rpm)/18°C/48 hrs or a vTi65 rotor at 400,000g (64,000 rpm)/18°C/5 hrs. Following the collection of the plasmid from the gradient, the salt concentration was reduced by the addition of 2.5 volumes of water. Nucleic acid was ethanol precipitated, washed, dried and resuspended in 1 ml of water. If further purification of the DNA was not required the ethidium bromide was solvent extracted from the aqueous solution using CsCl saturated isopropanol before precipitation. To remove any residual RNA, the plasmid was then carefully loaded on top of a 10-40% w/v sucrose gradient in a Beckman SW27 tube (500μl of plasmid solution per tube) and spun at 100,000g (20,000 rpm)/18°C/17 hrs. The ethidium bromide was then solvent extracted from the aqueous solution using isoamyl alcohol, and the plasmid DNA recovered as above.

DNA Transfection of Rat-1 Cells.

Essentially similar methods were used for short-term (transient expression) and long-

term (cell transformation) assays. For transient expression studies, Rat-1 cells were seeded at $5x10^6$ cells per bioassay tray 24 hours prior to transfection, and were refed 4 hours before transfection with 90ml DMEM, 0.1% glutamine, 5% FCS per tray. DNA for transfection was prepared as a calcium phosphate precipitate by mixing together equal volumes of two freshly prepared solutions. Solution A contained 2xHBS (1% w/v HEPES, 1.6% w/v NaCl, pH adjusted to 7.1 ±0.05 with NaOH) and 2xPO4 (0.7mM Na_2HPO_4, 0.7mM NaH_2PO_4). Solution B, formed by rapid mixing, contained 10-20μg/ml plasmid DNA in 250mM $CaCl_2$. In order to produce a fine precipitate solution B was added dropwise to solution A, whilst bubbling air through solution A with a Pasteur pipette. The resulting mixture was allowed to stand at RT for 30±10 mins, during which time a precipitate formed. The sample was mixed vigorously before being added dropwise to the medium (10 ml per tray). Cells were incubated overnight at 32°C, following which the medium was removed and the cells were washed twice with versene and re-fed with 100ml DMEM, 0.1% w/v glutamine, 5% v/v FCS. RNA was harvested from the cells 48 hrs later, and thoroughly treated with RNase free DNase I in order to remove traces of transfected plasmid DNA. For cell transformation assays, 300ng of mT gene equivalent was transfected onto $3x10^5$ cells in a 5 cm dish, using 0.5 ml of precipitate and Rat-1 DNA as carrier. After overnight incubation at 32°C, cells were trypsinised and split between five 5cm dishes. After 15 days growth at 37°C, with refeeding at 3 day intervals, foci were counted following Leis hman staining.

Chloramphenicol acetyl transferase (CAT) assays.

The method of Gorman *et al.* (1982) was used after transfection of Rat-1 cells with the appropriate plasmid. Cells were seeded the day prior to transfection at $3x10^5$ per 5 cm dish and were refed the following morning. Calcium phosphate precipitates were prepared as described above, and cells were harvested 48 hrs after transfection. All further treatments were carried out with cold solutions and the cells on ice. The cells were washed

three times with complete PBS, 1ml TEN (100 mM NaCl, 10 mM Tris pH 7.4, 1mM EDTA pH 7.4)was added per 5 cm dish and the cells were held on ice for 5 minutes. The cells were scraped off using a rubber policeman, transferred to eppendorf tubes and spun at 12,000g for 5 minutes. After carefully removing the supernatant by aspiration 50 µl of 250 mM Tris pH 7.8 was added to the sedimented cells. Each sample was then sonicated for 12-15 seconds using a MSE sonicator with a microprobe at 12 microns strength. Debris was removed by centrifugation at 12,000g for 5 minutes, and the supernatant was then used in the assay. ^{14}C-Chloramphenicol (CM) (50 µCi, 50 mCi/mM) was made up to a final volume of 1.6 ml by adding 0.67 volumes of 250 mM Tris pH 7.5, to give a final concentration of 100 mM Tris, with 0.375 µCi CM per 12 µl Each reaction mixture contained 10µl lysate, 10µl water, 12 µl CM/Tris and 5 µl AcCoA (4mM). Incubation was for 60 minutes at 37°C, after which time the reaction was stopped by the addition of 0.5 ml ethylacetate. Each tube was vortexed for 10 secs, and then spun at 12,000g for 1 min. to separate the phases. 0.45 ml of the upper (organic) phase was then transferred to a fresh tube, and dried down. Each sample was resuspended in 25 µl dyed ethylacetate before loading onto Macherey-Nagel Polygram SIL/UV_{254} precoated plastic sheets and ascending chromatography. Using these plates it was possible to assess the success of a reaction by viewing the products under a UV (254 nm) lamp. The solvent for TLC was a 95:5 mixture of chloroform:methanol. After drying, the plates were sprayed with EN^{3}HANCE and exposed to preflashed X-ray film. In order to ensure that the enzyme reaction was linear over the time of incubation, the positive control was assayed first, taking equal samples at 15 minute intervals.

Agarose Gel Electrophoresis of DNA.

Large submarine agarose gels (150-200ml) and mini submarine agarose gels (50ml) were made and run in 1xTBE. Depending on the size of the DNA fragments under investigation the agarose concentration varied from 0.7-2% w/v. All gels and running

buffers contained 0.5µg/ml ethidium bromide to enable visualisation of the DNA under long wave (360nm) ultraviolet light. Samples were loaded following the addition of 0.1x volume of loading buffer (25% w/v Ficoll, 50mM EDTA pH8, 0.1% w/v xylene cyanol, 0.1% w/v bromophenol blue). Gels were run at 20-120 volts for 1-16 hrs.

Acrylamide Gel Electrophoresis.

Stock solutions of acrylamide for sequencing gels, and non-denaturing preparative gels, contained 30:1 acrylamide:bisacrylamide, whilst those for strand separation ("cracking") gels were made with 66:1 stock solutions. All gels were polymerised using 250µl 10% w/v ammonium persulphate and 30µl of TEMED per 50ml gel volume, and pre-run before use. Sequencing gels and preparative sequencing gels were 0.4mm thick. These gels varied from 4-20% polyacrylamide in 1xTBE buffer containing 50% w/v urea, and were electrophoresed in 1xTBE at a current sufficient to maintain denaturing conditions within the gel (normally about 30 mA, 1.5kV). Samples obtained from dideoxy sequencing reactions were treated as outlined in Bankier and Barrell (1983) before loading. All other samples were resuspended in form stop buffer and heated to 90-100°C for 2 mins before loading. When run, gels containing ^{32}P radiochemicals were normally covered in Saran wrap and autoradiographed at -70°C (except preparative sequencing gels which were kept at RT), whilst dideoxy sequencing gels containing ^{35}S radiochemicals were fixed for 30 mins in a tray of 10% v/v methanol, 10% v/v acetic acid, and then dried onto 3MM paper at 80°C for 30 mins, before autoradiography (without a screen) at RT. Non-denaturing preparative gels were 0.8 mm thick and were typically 8% polyacrylamide in 1xTBE buffer. Samples were loaded in agarose gel loading buffer (see above) and the gels were run cold (300V) in 1xTBE. Strand separating ("cracking") gels were 0.8mm thick and were typically 5% polyacrylamide in 0.5xTBE buffer. Samples were denatured in form stop buffer, or strand separation buffer (50% v/v dimethyl sulphoxide, 1mM EDTA, 0.02% w/v bromophenol blue and xylene cyanol), by heating at 90-100°C before

loading, and the gels run cold (300V) in 0.5xTBE overnight.

Isolation of Specific DNA Fragments from Gels.

This method was used for the isolation of DNA fragments from both agarose and acrylamide gels. The excised section of the gel containing the required band was placed inside a dialysis bag with TE plus 100μg/ml BSA. This was then placed in a gel tank of 0.5-1x TBE and the DNA electroeluted at 100 volts for 1-2 hrs. The eluted DNA was extracted with sequentially with 2FC and chloroform, ethanol precipitated, washed, dried and resuspended in TE.

Southern blotting.

DNA in agarose gels was denatured in 500ml of GS1 for 1 hr, rinsed and then similarly neutralised with GS2. To assist transfer of high molecular weight DNA, gels were sometimes pre-soaked in 0.25M HCl for 20 mins before denaturation to partially acid hydrolyse the DNA. DNA was blotted onto nitrocellulose, pre-wet in 10xSSC, via capillary action from a reservoir containing 10x SSC as described in Maniatis et al. (1982). Following overnight transfer of DNA, the filters were air dried and baked under vacuum at 80°C for 2 hrs. Usually 15-20μg per track of genomic DNA, or 0.1-0.5μg of plasmid DNA, was used for Southern blotting and markers were provided by an end labelled *Hin*dIII restriction digest of λ DNA.

Nitrocellulose Filter Hybridisation.

After baking, filters were first prehybridised in 50-100ml of a solution containing 10x Denhardts, 6xSSC, 0.5% w/v SDS and 50μg/ml sonicated denatured salmon sperm DNA for approximately 2 hrs at 68°C (alternatively a prehybridisation solution containing 50% deionised formamide and a temperature of 45°C was used). After removal of the pre-hybridisation fluid, filters were hybridised in 5-10ml of a solution containing 2x Denhardts, 6xSSC, 0.5% w/v SDS and 50μg/ml sonicated denatured salmon sperm DNA

and 0.1μg denatured probe DNA at 68°C overnight (or alternatively, with 50% formamide at 45°C). Filters were washed at 68°C twice in 2xSSC, 0.5% w/v SDS and a third time in 0.1-0.5xSSC, 0.5% w/v SDS (depending on the stringency required), each wash being for 20-60 mins. Following washing, filters were air dried, covered with Saran wrap and autoradiographed. To re-probe Southern blots the old probe was first stripped off by soaking the filter in 50mM NaOH for 1 min. The filter was then rinsed in water and soaked in 0.1M Tris.Cl pH7.4 for 5 mins. Following a further rinse with water the filter was air dried and autoradiographed to ascertain if all the old probe had been removed before re-probing.

Labelling DNA to high specific activity with ^{32}P.

DNA hybridisation probes were made using supercoiled plasmid DNA, or purified DNA fragments, by nick translation (Rigby et al., 1977) or random oligonucleotide priming. For nick translation, typically 0.5μg of the DNA was labelled with 30μCi (α-^{32}P)-dCTP and 30μCi (α-^{32}P)-dATP (both 3,000 Ci/mmol) using *E.coli* DNA polymerase I in nick translation buffer (50mM Tris.Cl pH7.5, 5mM $MgCl_2$, 1mM β-mercaptoethanol, 50μg/ml BSA, 2μM dTTP and dGTP) for 1-3 hrs at 14°C. For nick translations involving whole cell DNA, 0.1ng/ml DNase I was also included in the reaction mixture to assist nicking of the DNA. Specific activities of $2x10^8$ dpm/μg of DNA were routinely obtainable. Unincorporated nucleotides were removed by column chromatography over G75 Sephadex using 0.1M NaCl, TE. The probe DNA represented the first column peak, and was either used directly or stored at -20°C. For random priming, DNA (50-200ng, 17μl) was denatured by boiling for 5 mins, and after briefly chilling on ice 5μl OLB (250mM Tris-HCl pH 8.0, 25 mM $MgCl_2$, 100μM each dTTP and dGTP, 1M HEPES pH6.6, 1mg/ml pd$(N)_6$ (sodium salt)), 1μl BSA (10 mg/ml), 2.5μl each (α-^{32}P)dC and dATP (300 Ci/mMol) and 2 units large fragment DNA polymerase I (Klenow polymerase) were added. Incubation was at room temperature, 3

hours to overnight. The labelled DNA was separated from unincorporated nucleotides as described above.

Formaldehyde Agarose Gel Electrophoresis of RNA and Northern Blotting.

The procedures used were modified from that presented by Maniatis *et al.* (1982). 200ml gels were cast which contained 1% w/v agarose in 1xMOPS-E, 6.8% w/v formaldehyde. Samples for electrophoresis, usually 1-2μg of poly(A)$^+$ RNA or 20μg of poly(A)$^-$ RNA, were prepared in 1xMOPS-E, 6.8% w/v formaldehyde, 50% v/v formamide and heated at 55°C for 15-30 mins following which 0.1x volume of RNA loading buffer (50% v/v glycerol, 1mM EDTA, 0.4% bromophenol blue, 0.4% xylene cyanol) was added per sample prior to loading on the gel. Gels were electrophoresed in 1xMOPS-E, 6.8% w/v formaldehyde at 30mA overnight with the buffer circulating by means of a peristaltic pump. Gels were run until the bromophenol blue had travelled 75-100% their length. Gels were soaked in water for 5 mins, followed by 50mM NaOH for 20-30 mins, to partially hydrolyse the RNA for ease of transfer during blotting. Gels were then neutralised by soaking in 0.1M Tris.Cl pH7.4 for 20-30 mins and finally soaked in 20xSSC for 60 mins. The gel was laid on a glass plate and the RNA transferred to nitrocellulose without using a reservoir. (The nitrocellulose was first pre-wet in water, and then placed in 20xSSC for 5 min, before being applied to the gel). After 3-12 hrs of RNA transfer, the nitrocellulose was air dried, without rinsing, and then baked under vacuum at 80°C for 2 hrs. Hybridisation and washing procedures were identical to those used for Southern blotting. After autoradiography the filter was stained for RNA to provide ribosomal RNA markers. The filter was soaked in 5% v/v acetic acid for 15 mins and then for in 0.5M NaOAc pH5.4, 0.04% w/v methylene blue for 5-10 mins. The filter was then rinsed with water to reveal the RNA. Alternatively, an RNA size marker ladder (BRL) was run in an isolated slot on the gel. The marker track was then cut off, blotted and stained as described above.

End-labelling DNA.

DNA for 5' end-labelling was dephosophorylated and free of contaminating RNA. 1μg of the DNA was labelled with 100μCi (γ-^{32}P)ATP (>5,000 Ci/mmol) using T4 polynucleotide kinase in kinase buffer (50mM TrisHCl pH7.5, 0.1mM EDTA, 10mM $MgCl_2$, 5mM DTT, 0.1mM spermidine) for 60 mins at 37°C. The volume was then increased to 200μl, containing 0.3M NaOAc pH 5.4, and the mixture extracted with 2FC, then chloroform and finally ethanol precipitated. Where the labelled DNA was to be used for size markers no further manipulation was necessary. However when the labelled DNA was to be used for sequencing, S1 mapping, or primer extension, the labelled end required had first to be purified away from the other labelled end. In most cases this was performed by cleaving with a second restriction enzyme that produced fragments that could be easily separated on agarose gels, or non-denaturing acrylamide preparative gels (see below). If no convenient second enzyme existed, or when single stranded sequences were required for primer extension analysis, the strands were separated on acrylamide "cracking" gels (see above). In order to ascertain which was the correct strand to be used, both strands from "cracking" gels were Maxam-Gilbert sequenced before use. Alternatively, the two strands could be distinguished by comparison of a sample labelled at both available ends with one labelled at only one (known) end. This was achieved by labelling the end(s) either before or after cutting with a second restriction enzyme. In order to label recessed 3' termini, plasmid digests or purified DNA fragments (0.2 to 1 μg) were incubated in 30μl of restriction enzyme buffer containing 20μCi of the appropriate α-labelled dNTP (3,000 Ci/mMol) and 1 unit of Klenow polymerase for 15 mins at room temperature. The volume was then increased to 200μl, containing 0.3M NaOAc pH 5.4, and the mixture extracted with 2FC, then chloroform and finally ethanol precipitated.

S1 Nuclease Mapping.

A typical reaction contained 20,000 cpm (Cerenkov) of probe DNA (approximately

5ng) mixed with either 40µg of total RNA or 2-5µg of poly $(A)^+$ RNA, whilst negative control reactions contained 40µg of yeast RNA. After mixing the RNA and DNA were freeze dried, and the lyophilised sample resuspended in 20µl of annealing buffer (40mM PIPES pH 6.4, 1mM EDTA pH8, 0.4M NaCl, 80% v/v formamide). Samples were denatured at 85-90°C for 15 mins, and then immediately transferred to a water bath, set at the required temperature, for overnight annealing. The temperature was set to maximise RNA:DNA hybridisation (as opposed to reannealing of double stranded DNA probes) from the equation determined by Thomas *et al.* (1976):

$$t_{max} = 81.5 + 0.5(\%G+C) + 16.6\ (\log[Na^+]) - 0.6(\%formamaide)$$

which, for the annealing buffer used, can be simplified to:

$$t_{max} = 26.9 + 0.5(\%G+C)$$

Thus the temperature of hybridisation varied between different experiments, depending on the (%G+C) content of the probe. In cases where more than one single-stranded probe was used simultaneously, the lowest calculated temperature was chosen. After overnight hybridisation, without allowing any prior cooling of the annealing mix, 300µl of ice cold S1 digestion buffer (280mM NaCl, 50mM NaOAc pH4.6, 4.5 mM $ZnSO_4$, 20µg/ml denatured salmon sperm DNA, 400 units/ml nuclease S1) was added to each sample, which was quickly mixed and placed on ice. When all the samples had been treated in this way, they were incubated at 37°C for 30 mins, or at 15°C for 1-2 hrs when shorter protected fragments were expected. Reactions were stopped by the addition of 75µl of S1 stop buffer (50mM EDTA, 2.5M NH_4OAc), 2µg of yeast RNA added and the samples precipitated on dry ice using 400µl of isopropanol. Washed, dried samples were then resuspended in 5µl of form stop buffer, heated to 90-100°C for 5 mins and loaded onto acrylamide sequencing gels or alkaline agarose gels, prepared and run as described by Maniatis *et al.* (1983).

Double-stranded probes were used in all cases, except where two probes were used simultaneously; in the latter case single-stranded probes were used.

cDNA Synthesis.

Oligo(dT)-primed first strand synthesis was initiated by adding 10µl oligo(dT) (1mg/ml), 5µl Tris-HCl (1M, pH 8.3 at 42°C), 5µl $MgCl_2$ (0.2M) and 14µl KCl (1M) to 10 µg 53-Rat poly $(A)^+$ RNA in 32µl. After heating to 65°C for 10 mins and briefly chilling on ice, the following were added: 1µl DTT (1M), 2.5µl RNAsin, 5µl each (α-^{32}P)dC and dATP (3,000 Ci/mMol), 2.5µl each dC and dATP (25mM), 5µl each dG and dTTP (25mM) and 5µl super RT (AMV reverse transcriptase). Reverse transcription was performed for 2 hours at 42°C, after which time unincorporated nucleotides were removed by Sephadex G-75 chromatography using 100mM NaCl, 0.1% SDS in TE. Labelling the cDNA to a relatively low specific activity in this way provided a useful means of detection of small quantities of product at later stages. A small aliquot of the reverse-transcribed product was run out on a sequencing-type gel in order to assess its average size. 1µg of dextran was added as carrier and the nucleic acid precipitated with ethanol following the addition of NH_4OAc to 2M. The samples were resuspended in 2M NH_4OAc, reprecipitated and washed with 70% v/v ethanol before further use. Second strand synthesis was performed by the method of Gubler and Hoffman (1983), in SSS buffer (20mM Tris.Cl pH 7.4, 5mM $MgCl_2$, 10mM $(NH_4)_2SO_4$, 100mM KCl, 0.15mM β-NAD, 40µM each dNTP, 50µg/ml BSA) using 8.5 units/ml *E.coli* RNase H, 230 units/ml DNA polymerase I and 10 units/ml *E.coli* ligase. Reactions were incubated at 12°C for 60 mins, followed by 22°C for 60 mins. Reactions were stopped by adding EDTA to 20mM and SDS to 0.1% w/v final. Samples were extracted twice with 2FC, once with chloroform and ethanol precipitated twice in the presence of 2M NH_4OAc. The ds cDNA was resuspended in T4 DP buffer (33mM Tris-acetate pH 7.9, 66mM KOAc, 10mM MgOAc, 0.5mM DTT, 0.1mg/ml BSA) and treated with T4 DNA polymerase for 5 mins at 37°C. Subsequently dNTPs were added to a final concentration of 200µM and the incubation continued at 37°C for a further 30 mins. Samples were extracted with 2FC, then chloroform and ethanol precipitated in the presence of 2M NH_4OAc. *Eco*RI linkers

were then ligated to the repaired cDNA, and excess linkers removed with *Eco*RI. cDNA was purified away from linkers by agarose gel electrophoresis and size-selected to be greater than 70 bp in length. The cDNA was then cloned into the *Eco*RI site of λgt10 and screened as described above. cDNA inserts of interest were all subsequently subcloned into M13mp18 or 19 and sequenced.

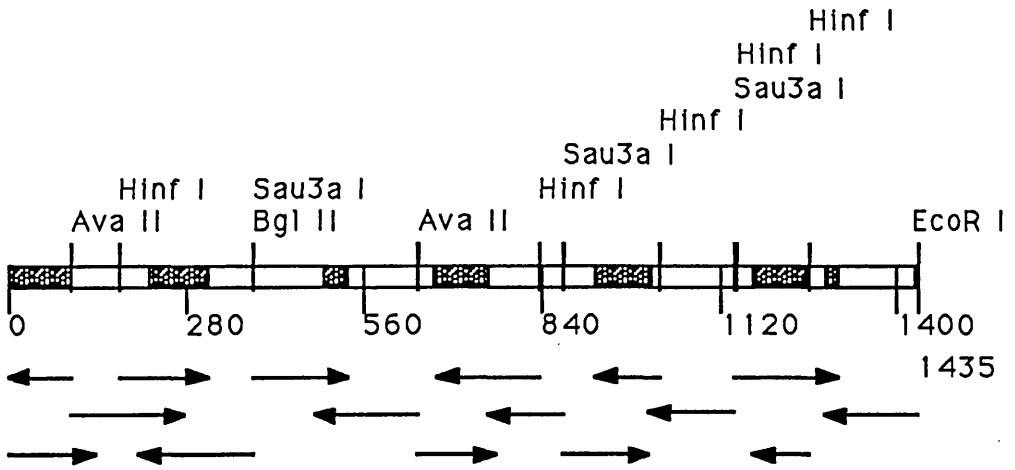
DNA Sequence Analysis.

The majority of DNA sequence analysis was performed by the dideoxy chain termination method (Sanger *et al.*, 1977) using recombinant M13 templates grown in *E.coli* JM101 as detailed in Bankier and Barrell (1983), with the modification that templates were protein extracted twice with TE saturated phenol, and once with chloroform, before use. End-labelled DNA was sequenced as described by Maxam and Gilbert (1977).

Computer Analysis of Sequence Data.

Primary sequence data were entered as DEC-20 files using GENED (IntelliGenetics) and were analysed using SEQ (Stanford University). Global nucleic acid homology searches were performed using both EMBL and GENBANK databases and the LSEARCH and SRCHE programs of G. Soundy (ICRF Research Computer Unit).

Sequencing strategy for 53-Rat cellular sequences.



M13 mp18 and mp19 subclones were sequenced in the directions indicated by arrows. The left-hand extent of the map shown corresponds to the viral-host junction: for sequencing and sub-cloning purposes the *Fnu*DII site which occurs at the junction was converted to an *Eco*RI and *Bam*HI site by linker addition. Where the sequence was confirmed on both complementary strands, this is indicated by shading of the map.

REFERENCES

Abramczuk, J., Pan, S., Maul, G., and Knowles, B. B. 1984. J. Virol. **49**, 540-548.

Acheson, N. H. 1978. Proc. Natl. Acad. Sci. USA **75**, 4754-4758.

Acheson, N. H. 1984. Mol. Cell. Biol. **4**, 722-729.

Acheson, N. H., and Mieville, F. 1978. J. Virol. **28**, 885-894.

Adhya, S., and Gottesman, M. 1982. Cell **29**, 939-944.

Alt, F. W., Bothwell, A. L. M., Knapp, M., Siden, E., Mather, E., Koshland, M., and Baltimore, D. 1980. Cell **20**, 293-301.

Alwine, J. C. 1982. J. Virol. **43**, 987-996.

Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S., and Evans, R. M. 1982. Nature **298**, 240-244.

Augerau, P., and Wasylyk, B. 1984. Nucl. Acids Res. **12**, 8801-8818.

Bacheler, L. T. 1977. J. Virol. **22**, 54-64.

Banerji, J., Rusconi, S., and Schaffner, W. 1981. Cell **27**, 299-308.

Bankier, A. T., and Barrell, B. G. 1983. In: Techniques in The Life Sciences. Flavell, R. A. (ed.). **B508**, 1-34.

Basilico, C., Gattoni, S., Zouzias, D., and Della Valle, G. 1979a. Cell **17**, 645-659.

Basilico, C., Zouzias, D., Della Valle, G., Gattoni, S., Colantuoni, V., Fenton, R., and Dailey, L. 1980. Cold Spring Harbor Symp. Quant. Biol. **44**, 611-620.

Baty, D., Barrera-Saldana, H. A., Everett, R. D., Vigneron, M., and Chambon, P. 1984. Nucl. Acids Res. **12**, 915-929.

Been, M.D., Burgess, R.R., and Champoux, J.J. 1984. Nucl. Acids Res. **12**, 3097.

Ben-Asher, E., and Aloni, Y. 1984. J. Virol **52**, 266-276.

Bender, M. A., and Brockman, W. W. 1981. J. Virol. **38**, 872-879.

Bendig, M. M., Thomas, T., and Folk, W. R. 1980. Cell **20**, 401-409.

Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. 1980. Nucl. Acids Res. **8**, 127-142.

Benoist, C., and Chambon, P. 1981. Nature **290**, 304-315.

Bentley, D. L., and Groudine, M. 1986. Nature **321**, 702-706.

Berg, P. E., Yu, J-K, Popovic, Z., Schumperli, D., Johansen, H., Rosenberg, M., and Anderson, W. F. 1983. Mol. Cell. Biol. **3**, 1246-1254.

Berg, P. E., and Anderson, W. F. 1984. Mol. Cell. Biol. **4**, 368-370.

Berger, H., and Wintersberger, E. 1986. J. Virol. **60**, 768-770.

Berget, S. M. 1984. Nature **309**, 179-181.

Bhat, B. M., and Wold, W. S. M. 1985. Mol. Cell. Biol. **5**, 3183-3193.

Bikel, I., Mamon, H., Brown, E. L., Boltax, J., Agha, M., and Livingston, D. 1986. Mol. Cell. Biol. **6**, 1172-1182.

Birchmeier, C., Schumperli, D., Sconzo, G., and Birnstiel, M. L. 1984. Proc. Natl. Acad. Sci. USA **81**, 1057-1061.

Birg, F., Favaloro, J. M., and Kamen, R. I. 1977. Proc. Natl. Acad. Sci. USA **74**, 3138-3142.

Birg, F., Dulbecco, R., Fried, M., and Kamen, R. 1979. J. Virol. **29**, 633-648.

Birnstiel, M. L., Busslinger, M., and Strub, K. 1985. Cell **45**, 349-359.

Bjursell, G. 1978. J. Virol. **26**, 136-142.

Blattner, F. R., Blechl, A., Denniston-Thompson, K., Faber, H., Richards, J. E., Slightom, J. L., Tucker, P., and Smithies, O. 1978. Science **202**, 1279-1284.

Bolen, J. B., Anders, D. G., Trempy, J., and Consigli, R. A. 1981. J. Virol. **37**, 80-91.

Botchan, M., Topp, W., and Sambrook, J. 1976. Cell **9**, 269-287.

Botchan, M., Stringer, J., Mitchison, T., and Sambrook, J. 1980. Cell **20**, 143-152.

Bradley, M. K., Griffin, J. D., and Livingston, D. M. 1982. Cell **28**, 125-134.

Bradley, M. K., Hudson, J., Villanueva, M. S., and Livingston, D. M. 1984. Proc. Natl. Acad. Sci. USA **81**, 6574-6578.

Brady, J. N., Winston, V. D., and Consigli, R. A. 1978. J. Virol. **27**, 193-204.

Brady, J. N., Bolen, J. B., Radonovich, M., Salzman, N., and Khoury, G. 1984. Proc. Natl. Acad. Sci. USA **81**, 2040-2044.

Brady, J. N., Loeken, M. R., and Khoury, G. 1985. Proc. Natl. Acad. Sci. USA **82**, 7299-7303.

Breathnach, R., and Chambon, P. 1981. Ann. Rev. Biochem. **50**, 349-383.

Breitbart, R. E., Nguyen, H. T., Medford, R. M., Destree, A. T., Mahdavi, V., and Nadal-Ginard, B. 1985. Cell **41**, 67-82.

Breitbart, R. E., and Nadal-Ginard, B. 1987. Cell **49**, 793-803.

Brown, P.O., Bowerman, B., Varmus, H.E., and Bishop, J.M. 1987. Cell **49**, 347-356.

Brugge, J. S., and Butel, J. S. 1975. J. Virol. **15**, 619-635.

Buchman, A. R., and Berg, P. 1984. Mol. Cell. Biol. **4**, 1915-1928.

Bullock, P., Champoux, J. J., and Botchan, M. R. 1985. Science **230**, 954-958.

Byrne, B. J., Davis, M. S., Yamaguchi, J., Bergsma, D. J., and Subramanian, K. N. 1983. Proc. Natl. Acad. Sci. USA **80**, 721-725.

Campbell, B. A., and Villarreal, L. P. 1985. Mol. Cell. Biol. **5**, 1534-1537.

Campo, M. S., Cameron, I. R., and Rogers, M. E. 1978. Cell **15**, 1411-1426.

Canaani, D., Kahana, C., Mukamel, A., and Groner, Y. 1979. Proc. Natl. Acad. Sci. USA **76**, 3078-3082.

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H., and Goeddel, D. V. 1983. Nature **302**, 33-34.

Champoux, J. J., and Been, M. D. 1980. In: Mechanistic Studies of DNA Replication and Genetic Recombination (Alberts, B., ed.). pp 809-924. Academic Press, New

York.

Chandrasekharappa, S. C., Smith, J. H., and Elceiri, G. L. 1983. J. Cell. Physiol. **117**, 169-174.

Chen-Kiang, S., Wolgemuth, D. J., Hsu, M.-T., and Darnell, Jr., J. E. 1982. Cell **28**, 575-584.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. 1979. Biochemistry **18**, 5295-5299.

Citron, B., Falck-Pedersen, E., Salditt-Georgieff, M., and Darnell, Jr., J. E. 1984. Nucl. Acids Res. **12**, 8723-8731.

Clark, R., Trevethia, M. J., and Tjian, R. 1984. Cancer cells **2**, 363-368.

Cogen, B. 1978. Virology **85**, 222-230.

Cole, C. N., and Stacy, T. P. 1985. Mol. Cell. Biol. **5**, 2104-2113.

Cole, C. N., Tornow, J., Clark, R., and Tjian, R. 1986. J. Virol. **57**, 539-546.

Courtneidge, S. A., and Smith, A. E. 1983. Nature **303**, 435-439.

Courtneidge, S. A., and Smith, A. E. 1984. EMBO J. **3**, 585-591.

Cowan, K., Tegtmeyer, P., and Anthony, D. D. 1973. Proc. Natl. Acad. Sci. USA **70**, 1927-1930.

Cowie, A., Tyndall, C., and Kamen, R. 1981. Nucl. Acids Res. **9**, 6305-6322.

Cowie, A., Jat, P., and Kamen, R. 1982. J. Mol. Biol. **159**, 225-255.

Cowie, A., and Kamen, R. 1984. J. Virol. **52**, 750-760.

Dailey, L., and Basilico, C. 1985. J. Virol. **54**, 739-749.

Danner, D., and Leder, P. 1985. Proc. Natl. Acad. Sci. USA **82**, 8658-8662.

Davidson, D., and Hassell, J.A. 1987. J. Virol. **61**, 1226-1239.

Davison, B. L., Egly, J., Mulvihill, E. R., and Chambon, P. 1983. Nature **301**, 680-686.

Defendi, V.1960. Nature **188**, 508-509.

De Ferra, F., Engh, H., Hudson, L., Kamholz, J., Puckett, C., Molineaux, S., and

Lazzarini, R. A. 1985. Cell **43**, 721-727.

Deininger, P., Esty, A., LaPorte, P., and Friedmann, T. 1979. Cell **18**, 771-779.

Deininger, P. L., Esty, A., LaPorte, P., Hsu, H., and Friedmann, T. 1980. Nucl. Acids Res. **8**, 855-860.

Della Valle, G., Fenton, R. G., and Basilico, C. 1981. Cell **23**, 347-355.

Delmas, V., Bastien, C., Scherneck, S., and Feunteun, J. 1985. EMBO J. **4**, 1279-1286.

DeNoto, F. M., Moore, D. D., and Goodman, H. M. 1981. Nucl. Acids Res. **9**, 3719-3730.

de Villiers, J., and Schaffner, W. 1981. Nucleic Acids Res. **9**, 6251-6264.

de Villiers, J., Olsen, L., Tyndall, C., and Schaffner, W. 1982. Nucl. Acids Res. **10**, 7965-7976.

de Villiers, J., Schaffner, W. Tyndall, C., Lupton, S., and Kamen,R. 1984. Nature **312**, 242-246.

Diamandopoulos, G. T. 1978. J. Natl. Cancer Inst. **60**, 445-450.

Dilworth, S. M., Cowie, A., Kamen, R. I., and Griffin, B. E. 1984. Proc. Natl. Acad. Sci. USA **81**, 1941-1945.

Di Mayorca, G., Callender, J., Marin, G., and Giordano, R. 1969. Virology **38**, 126-133.

Donehower, L., Huang, A., and Hager, G. 1981. J. Virol. **37**, 226-238.

Dressler, G.R., and Fraser, N.W. 1987. J. Virol. **61**, 2770-2776.

Dulbecco, R., and Eckhart, W. 1970. Proc. Natl. Acad. Sci. USA **67**, 1775-1781.

Dynan, W. S., and Tjian, R. 1983a. Cell **32**, 669-680.

Dynan, W. S., and Tjian, R. 1983b. Cell **35**, 79-87.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. 1980. Cell **20**, 313-319.

Eckhart, W., Hutchinson, M. A., and Hunter, T. 1979. Cell **18**, 925-934.

Eddy, B. E., Borman, G. S., Grubbs, G. E., and Young, R. D. 1962. Virology **17**, 65-75.

Ernoult-Lange, M., May, P., Moreau, P,. and May, E. 1984. J. Virol. **50**, 163-173.

Everett, R. D., Baty, D., and Chambon, P. 1983. Nucl. Acids Res. **11**, 2447-2464.

Falck-Pedersen, E., Logan, J., Shenk, T., and Darnell, Jr., J. E. 1985. Cell **40**, 897-905.

Fenton, R. G., and Basilico, C. 1981. J. Virol. **40**, 150-163.

Fenton, R. G., and Basilico, C. 1982. Virology **121**, 384-392.

Ferdinand, F.-J., Brown, M., and Khoury, G. 1977. Virology **78**, 150-161.

Feunteun, J., Sompayrac, L., Fluck, M., and Benjamin, T. 1976. Proc. Natl. Acad. Sci. USA **73**, 4169-4173.

Fitzgerald, M., and Shenk, T. 1981. Cell **24**, 251-260.

Flavell, A. J., and Kamen, R. I. 1977. J. Mol. Biol. **115**, 237-242.

Floros, J., Jonak, G., Galanti, N., and Baserga, R. 1981. Exp. Cell Res. **132**, 215-223.

Ford, M., Davies, B., Griffiths, M., Wilson, J., and Fried, M. 1985. Proc. Natl. Acad. Sci. USA **82**, 3370-3374.

Ford, J. P., and Hsu, M.-T. 1978. J. Virol. **28**, 795-801.

Francke, B., and Eckhart, W. 1973. Virology **55**, 127-137.

Frayne, E.G., Leys, E.J., Crouse, G.F., Hook, A.G., and Kellems, R.E. 1984. Mol. Cell. Biol. **4**, 2921-2924.

Freed, M. I., Lubin, I., and Simmonds, D. T. 1983. J. Virol. **46**, 1061-1065.

Fried, M. 1965. Proc. Natl. Acad. Sci. USA **53**, 486-491.

Fried, M., Griffiths, M., Davies, B., Bjursell, G., LaMantia, G., and Lania, L. 1983. Proc. Natl. Acad. Sci. USA **80**, 2117-2121.

Fried, M., Wilson, J. B., Williams, T., and Norbury, C. 1986. Cancer Cells **4**, 309-316.

Frischauf, A., Lehrach, H., Poustka, A., and Murray, N. 1983. J. Mol. Biol. **170**, 827-842.

Fromm, M., and Berg, P. 1982. J. Mol. Appl. Genet. **1**, 457-481.

Fromm, M., and Berg, P. 1983. Mol. Cell. Biol. **3**, 991-999.

Fujimura, F. K., Deininger, P. L., Friedmann, T., and Linney, E. 1981. Cell **23**, 809-814.

Garber, E. A., Seidman, M. M., and Levine, A. J. 1978. Virology **90**, 305-316.

Garcea, R. L., Ballmer-Hofer, K., and Benjamin, T. L. 1985. J. Virol. **54**, 311-316.

Gaudray, P., Tyndall, C., Kamen, R., and Cuzin, F. 1981. Nucl. Acids Res. **9**, 5697-5710.

Gerard, R. D., Woodworth-Gutai, M., and Scott, W. A. 1982. Mol. Cell. Biol. **2**, 782-788.

Ghosh, P. K., and Lebowitz, P. 1981. J. Virol. **40**, 224-240.

Gidoni, D., Kahana, C., Canaani, D., and Groner, Y. 1981. Proc. Natl. Acad. Sci. USA **78**, 2174-2178.

Gidoni, D., Dynan, W. S., and Tjian, R. 1984. Nature **312**, 409-413.

Gil, A., and Proudfoot, N. J. 1984. Nature **312**, 473-474.

Gil, A., and Proudfoot, N. J. 1987. Cell **49**, 399-406.

Girardi, A. J., Sweet, B. H., Slotnick, V. B., and Hilleman, M. R. 1962. Proc. Soc. Exp. Biol. Med. **109**, 649-660.

Gluzman, Y., and Shenk, T. (eds.). 1983. Enhancers and Eukaryotic Gene Expression. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Gorman, C., Moffat, L., and Howard, B. 1982. Mol. Cell. Biol. **2**, 1044-1051.

Graessman, A., Graesssman, M., and Mueller, C. 1981. Adv. Cancer Res. **35**, 111-149.

Graffi, A., Schramm, T., Bender, E., Bierwolf, D., and Graffi, I. 1967. Arch. Geschwulstforsch. **30**, 277-283.

Grass, D. S., Jove, R., and Manley, J. L. 1987. Nucl. Acids Res. **15**, 4417-4436.

Green, M. R. Ann. Rev. Genet. **20**, 671-708.

Greenspan, D. S., and Carroll, R. B. 1981. Proc. Natl. Acad. Sci. USA **78**, 105-109.

Griffin, B. E., and Fried, M. 1975. Nature **256**, 175-179.

Griffin, B. E., Soeda, E., Barrell, B. G., and Staden, R. 1981. In J. Tooze (Ed.). DNA Tumour Viruses (2nd. ed. revised), pp. 843-910. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Gross, L. 1953. Proc. Soc. Exp. Biol. Med. **83**, 414-421.

Grummt, I., Kuhn, A., Bartsch, I., and Rosenbauer, H. 1986. Cell **47**, 901-911.

Haegeman, G., and Fiers, W. 1978. J. Virol. **25**, 824-830.

Haegeman, G., and Fiers, W. 1980. J. Virol. **35**, 955-961.

Hanahan, D. 1983. J. Mol. Biol. **166**, 557-580.

Harland, R., and Weintraub, H. 1985. J. Cell Biol. **101**, 1094-1099.

Hart, R. P., McDevitt, M. A., and Nevins, J. R. 1985a. Cell **43**, 667-683.

Hart, R. P., McDevitt, M. A., Ali, H., and Nevins, J. R. 1985b. Mol. Cell. Biol. **5**, 2975-2983.

Hashimoto, C., and Steitz, J. A. 1986. Cell **45**, 581-591.

Hasson, J. F., Mougneau, E., Cuzin, F., and Yaniv, M. 1984. J. Mol. Biol. **177**, 53-68.

Hay, N., Skolnik-David, H., and Aloni, Y. 1982. Cell **29**, 183-193.

Hay, N., and Aloni, Y. 1984. Nucl. Acids Res. **12**, 1401-1414.

Hayday, A.C. 1982. Ph.D. Thesis, University of London.

Hayday, A. C., Ruley, H. E., and Fried, M. 1982. J. Virol. **44**, 67-77.

Hayday, A. C., Chaudry, F., and Fried, M. 1983. J. Virol. **45**, 693-699.

Heiser, W. C., and Eckhart, W. 1982. J. Virol. **44**, 175-188.

Helfman, D. M., Cheley, S., Kuismanen, E., Finn, L. A., and Yamawaki-Kataoka, Y. 1986. Mol. Cell. Biol. **6**, 3582-3595.

Henderson, S., and Sollner-Webb, B. 1986. Cell **47**, 891-900.

Henikoff, S., Sloan, J. S., and Kelley, J. D. 1983. Cell **34**, 405-414.

Herbomel, P., Saragosti, S., Blangy, D., and Yaniv, M. 1981. Cell **25**, 651-658.

Herbomel, P., Bourachot, B., and Yaniv, M. 1984. Cell **39**, 653-662.

Hernandez, N. 1985. EMBO J. **4**, 1827-1837.

Hernandez, N., and Weiner, A. M. 1986. Cell **47**, 249-258.

Herr, W., and Gluzman, Y. 1985. Nature **313**, 711-714.

Higgs, D. R., Goodbourn, S. E. Y., Lamb, J., Clegg, J. B., Weatherall, D. J., and Proudfoot, N. J. 1983. Nature **306**, 398-400.

Hiscott, J. B., Murphy, D., and Defendi, V. 1981. Proc. Natl. Acad. Sci. USA **78**, 1736-1740.

Holmes, D. S., and Quigley, M. 1981. Anal. Biochem. **114**, 193-197.

Horowitz, H., and Platt, T. 1982. Nucl. Acids Res. **10**, 5447-5465.

Huynh, T. V., Young, R. A., and Davis, R. W. 1985. In: DNA Cloning: a Practical Approach. Glover, D. M. (ed.). IRL Press, Oxford, England. **1**, 49-78.

Ito, Y., and Spurr, N. 1979. Cold Spring Harbor Symp. Quant. Biol. **44**, 149-157.

Ito, Y., Brocklehurst, J. R., and Dulbecco, R. 1977. Proc. Natl. Acad. Sci. USA **74**, 4666-4670.

Jacobovits, E. B., Bratosin, S., and Aloni, Y. 1982. Virology **120**, 340-348.

Jat, P., Novak, U., Cowie, A., Tyndall, C., and Kamen, R. 1982. Mol. Cell. Biol. **2**, 737-751.

Jat, P., Roberts, J. W., Cowie, A., and Kamen, R. 1982b. Nucl. Acids Res. **10**, 871-887.

Jay, G., Nomura, S., Anderson, C. W., and Khoury, G. 1981. Nature **291**, 346-349.

Johnson, M. R., Norman, C.,Reeve, M. A., Scully, J., and Proudfoot, N. J. 1986. Mol. Cell. Biol. **6**, 4008-4018.

Jones, K. A., and Tjian, R. 1984. Cell **36**, 155-162.

Jongstra, J., Reudelhuber, T. L., Oudet, P., Benoist, C., Chae, C.-B., Jeltsch, J.-M., Mathis, D. J., and Chambon, P. 1984. Nature **307**, 708-714.

Kahana, C., Gidoni, D., Canaani, D.,and Groner, Y. 1981. J. Virol. **37**, 7-16.

Kalderon, D., Richardson, W. D., Markham, A. F., and Smith, A. E. 1984a. Nature **311**, 33-38.

Kalderon, D., Roberts, B. L., Richardson, W. D., and Smith, A. E. 1984b. Cell **39**, 499-509.

Kamen, R., Lindstrom, D. M., Shure, H., and Old, R. W. 1975. Cold Spring Harbor Symp. Quant. Biol. **39**, 187-198.

Kamen, R., Favaloro, J., Parker, J., Treisman, R., Lania, L., Fried, M., and Mellor, A. 1979. Cold Spring Harbor Symp. Quant. Biol. **44**, 63-75.

Kamen, R., Jat, P., Treisman, R., and Favaloro, J. 1982. J. Mol. Biol. **159**, 189-224.

Katinka, M., and Yaniv, M. 1983. J. Virol. **47**, 244-248.

Kaufman, R. J., and Sharp, P. A. 1983. Mol. Cell. Biol. **3**, 1598-1608.

Keller, J. M., and Alwine, J. C. 1984. Cell **26**, 381-390.

Khandjian, E. W., and Turler, H. 1983. Mol. Cell. Biol. **3**, 1-8.

Khoury, G., and May, E. 1977. J. Virol. **23**, 167-176.

Kim, S. K., and Wold, B. 1985. Cell **42**, 129-138.

Kimura, G., and Itagaki, A. 1975. Proc. Natl. Acad. Sci. USA **72**, 673-677.

Klug, A. 1965. J. Mol. Biol. **11**, 424-431.

Konarska, M. M., Padgett, R. A., and Sharp, P. A. 1985. Cell **42**, 165-171.

Kozak, M. 1983. Cell **34**, 971-978.

Krause, M., and Hirsh, D. 1987. Cell **49**, 753-761.

Krieg, P. A., and Melton, D. A. 1984. Nature **308**, 203-206.

Kriegler, M., Perez, C. F., Hardy, C., and Botchan, M. 1984. Cell **38**, 483-491.

Kuhne, T., Wieringa, B., Reiser, J., and Weissman, C. 1983. EMBO J. **2**, 727-733.

Lai, E. C., Stein, J. P., Catterall, J. F., Woo, S. L. C., Mace, M. L., Means, A. R., and O'Malley, B. W. 1979. Cell **18**, 829-842.

Lamb, O., and Aloni, Y. 1975. J. Virol. **16**, 1171-1183.

Lane, D. P., and Crawford, L. V. 1979. Nature **278**, 261-263.

Lanford, R. E., and Butel, J. S. 1980a. Virology **105**, 303-313.

Lanford, R. E., and Butel, J. S. 1980b. Virology **105**, 314-327.

Lanford, R. E., and Butel, J. S. 1984. Cell **37**, 801-813.

Lang, K. M., and Spritz, R. A. 1983. Science **220**, 1351-1355.

Langford, C. J., Klinz, F.-J., Donath, C., and Gallwitz, D. 1984. Cell **36**, 645-653.

Lania, L., Griffiths, M., Cooke, B., Ito, Y., and Fried, M. 1980. Cell **18**, 793-802.

Lania, L., Hayday, A., and Fried, M. 1981. J. Virol. **39**, 422-431.

Learned, R. M., Smale, S. T., Haltiner, M. M., and Tjian, R. 1983. Proc. Natl. Acad. Sci. USA **80**, 3558-3562.

Lee, W., Haslinger, A., Karin, M., and Tjian, R. 1987. Nature **325**, 368-372.

Legon, S., Flavell, A. J., Cowie, A., and Kamen, R. 1979. Cell **16**, 373-378.

LeMeur, M. A., Galliot, B., and Gerlinger, P. 1984. EMBO J. **3**, 2779-2786.

Lev, Z., Kamen, R., and Manor, H. 1979. Virology **93**, 445-457.

Levine, A. J. 1983. Curr. Top. Microbiol. Immunol. **101**, 1-30.

Levinson, B., Khoury, G., Vande Woude, G., and Gruss, P. 1982. Nature **295**, 568-572.

Lewin, B. 1980. Cell **22**, 324-326.

Lewis, E. D., and Manley, J. L. 1986. Proc. Natl. Acad. Sci. USA **83**, 8555-8559.

Li, J. J., and Kelley, T. J. 1984. Proc. Natl. Acad. Sci. USA **81**, 6973-6977.

Lloubes, R., Baty, D., and Lazdunski, C. 1986. Nucl. Acids Res. **14**, 2621-2636.

Loenen, W. A., and Brammar, W. J. 1980. Gene **10**, 249-259.

Luciw, P. A., Bishop, J. M., Varmus, H. E., and Capecchi, M. R. 1983. Cell **33**,

705-716.

Luthman, H., Nilsson, M.-G., and Magnusson, G. 1982. J. Mol. Biol. **161**, 533-550.

Lycan, D. E., and Danna, K. J. 1983. J. Virol. **45**, 264-274.

Mackay, R. L., and Consigli. R. A. 1976. J. Virol. **19**, 620-636.

Maderious, A., and Chen-Kiang, S. 1984. Proc. Natl. Acad. Sci. USA **81**, 5931-5935.

Majello, B., La Mantia, G., Simeone, A., Boncinelli, E. and Lania, L. 1985. Nature **314**, 457-459.

Maniatis, T., Fritsch, E., and Sambrook, J. 1982. Molecular Cloning, a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Manley, J. L. 1983. Cell **33**, 595-605.

Manteuil, S., and Girard, M. 1974. Virology **60**, 438-454.

Martin, M. A., Khoury, G., and Fareed, G. C. 1975. Cold Spring Harbor Symp. Quant. Biol. **39**, 129-142.

Martin, R. G., and Chou, J. Y. 1975. J. Virol. **15**, 599-612.

Maruyama, K., and Oda, K. 1984. J. Virol. **49**, 479-489.

Mason, P. J., Elkington, J. A., Lloyd, M. M., Jones, M. B., and Williams, J. G. Cell **46**, 263-270.

Mather, E. L., Nelson, K. J., Haimovich, J., and Perry, R. 1984. Cell **36**, 329-338.

Maurer, R. A., Erwin, C. R., and Donelson, J. E. 1981. J. Biol. Chem. **256**, 10524-10528.

Maxam, A. M., and Gilbert, M. 1980. Meth. Enzymol. **65**, 499-560.

McCance, D. J. 1981. J. Virol. **39**, 958-962.

McDevitt, M. A., Imperiale, M. J., Ali, H., and Nevins, J. R. 1984. Cell **37**, 993-999.

McDevitt, M. A., Hart, R. P., Wong, W. W., and Nevins, J. R. 1986. EMBO J. **5**,

2907-2913.

McLauchlan, J., Gaffney, D., Whitton, J. L., and Clements, J. B. 1985. Nucl. Acids Res. **13**, 1347-1368.

McStay, B., and Reeder, R. H. 1986. Cell **47**, 913-920.

Melin, F., Pinon, H., Reiss, C., Kress, C., Montreau, N., and Blangy, D. 1985a. EMBO J. **4**, 1799-1803.

Melin, F., Pinon, H., Kress, C., and Blangy, D. 1985b. J. Virol. **53**, 862-866.

Melnick, J. L., Allison, A. C., Butel, J. S., Eckhart, W., Eddy, B. E., Kit, S., Levine, A. J., Miles, J. A. R., Pagano, J. S., Sachs, L., and Vonka, V. 1974. Intervirology **3**, 106-120.

Melton, D. A. 1985. Proc. Natl. Acad. Sci. **82**, 144-148.

Montell, C., Fisher, E. F., Caruthers, M. H., and Berk, A. J. 1983. Nature **305**, 600-605.

Moore, C. L., and Sharp, P. A. 1984. Cell **36**, 581-591.

Moore, C. L., and Sharp, P. A. 1985. Cell **41**, 845-855.

Moore, J. L., Chowdhury, K., Martin, M. A., and Isreal, M. A. 1980. Proc. Natl. Acad. Sci. USA **77**, 1336-1340.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M. P., and Chambon, P. 1981. Nucleic Acids Res. **9**, 6047-6068.

Mount, S. M. 1982. Nucl. Acids Res. **10**, 459-472.

Mueller, C. R., Mes-Masson, A., Bouvier, M., and Hassell, J. A. 1984. Mol. Cell. Biol. **4**, 2594-2609.

Muller, H., and Nitschke, R. 1986. Med. Microbiol. Immunol. **175**, 1-13.

Muller, W. J., Mueller, C. R., Mes, A.-M., and Hassell, J. A. 1983. J. Virol. **47**, 586-599.

Murakami, Y., Eki, T., Yamada, M., Prives, C., and Hurwitz, J. 1986a. Proc. Natl. Acad. Sci. USA **83**, 6347-6351.

Murakami, Y., Wobbe, C. R., Weissbach, L., Dean, F. B., and Hurwitz, J. 1986b.

Proc. Natl. Acad. Sci. USA **83**, 2869-2873.

Myers, R. M., Rio, D. C., Robbins, A. K., and Tjian, R. 1981. Cell **25**, 373-384.

Nagata, S., Tsuchiya, M., Asano, S., Yamamoto, O., Hirata, Y., Kubota, N., Oheda, M., Nomura, H., and Yamazaki, T. 1986. EMBO J. **5**, 575-581.

Nevins, J. R., and Darnell, Jr., J. E. 1978. Cell **15**, 1477-1493.

Nevins, J. R. 1983. Ann. Rev. Biochem. **52**, 441-466.

Ng, S.-C., Mertz, J. E., Sanden-Will, S., and Bina, M. 1985. J. Biol. Chem. **260**, 1127-1132.

Nilsson, S. V., and Magnusson, G. 1984. J. Virol. **51**, 768-775.

Norkin, L. C. 1977. J. Virol. **21**, 872-879.

Omilli, F., Lange, M., Borde, J., and May, E. 1986. Mol. Cell. Biol. **6**, 1875-1885.

Oren, M., Winocour, E., and Prives, C. 1980. Proc. Natl. Acad. Sci. USA **77**, 220-224.

Osborn, M., and Weber, K. 1975. J. Virol. **15**, 636-644.

Padgett, B. L., Walker, D. L., ZuRhein, G. M., Eckroade, R. J., and Dessel, B. H. 1971. Lancet **1**, 1257-1260.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. 1986. Ann. Rev. Biochem. **55**, 1119-1150.

Pannuti, A., La Mantia, G., and Lania, L. 1987. Nucl. Acids Res. **15**, 1595-1613.

Parker, C. S., and Topol, J. 1984. Cell **36**, 357-369.

Paucha, E., Mellor, A., Harvey, R., Smith, A. E., Hewick, R. M., and Waterfield, M. D. 1978. Proc. Natl. Acad. Sci. USA **75**, 2165-2169.

Piette, J., Kryszke, M., and Yaniv, M. 1985. EMBO J. **4**, 2675-2685.

Platt, T. 1986. Ann. Rev. Biochem. **55**, 339-372.

Pomerantz, B. J., Mueller, C. R., and Hassell, J. A. 1983. J. Virol. **47**, 600-610.

Pomerantz, B. J., and Hassell, J. A. 1984. J. Virol. **49**, 925-937.

Postel, E. H., and Levine, A. J. 1976. Virology **73**, 206-216.

Proudfoot, N. J., and Brownlee, G. G. 1976. Nature **263**, 211-2214.

Proudfoot, N. J. 1986. Nature **322**, 562-565.

Rayment, I., Baker, T. S., Caspar, D. L. D., and Murakami, W. T. 1982. Nature **295**, 110-115.

Reddy, V. B., Ghosh, P. K., Lebowitz, P., and Weissman, S. M. 1978. Nucl. Acids Res. **5**, 4195-4213.

Reddy, V. B., Ghosh, P. K., Lebowitz, P., Piatak, M., and Weissman, S. M. 1979. J. Virol. **30**, 279-296.

Reed, R., and Maniatis, T. 1986. Cell **46**, 681-690.

Reed, S. I., Stark, G. R., and Alwine, J. C. 1976. Proc. Natl. Acad. Sci. USA **73**, 3083-3087.

Rio, D. C., and Tjian, R. 1983. Cell **32**, 1227-1240.

Rio, D. C., and Tjian, R. 1984. J. Mol. Appl. Genet. **2**, 423-432.

Robbins, P. D., Rio, D. C., and Botchan, M. R. 1986. Mol. Cell. Biol. **6**, 1283-1295.

Roth, J. R., and Wilson, J. H. 1986. Mol. Cell. Biol. **6**, 4295-4304.

Ruley, H. E., Lania, L., Chaudry, F., and Fried, M. 1982. Nucleic Acids Res. **10**, 4515-4524.

Ruley, H.E., Chaudry, F., Read, M., Lania, L., and Fried, M. 1987. J. Gen. Virol. (in press).

Ruley, H.E., and Fried, M. 1983. Nature **304**, 181-184.

Rundell, K. 1982. J. Virol. **42**, 1135-1137.

Ruskin, B., and Green, M. R. 1985. Cell **43**, 131-142.

Sadofsky, M., and Alwine, J. C. 1984. Mol. Cell. Biol. **4**, 1460-1468.

Salunke, D. M., Caspar, D. L. D., and Garcea, R. L. 1986. Cell **46**, 895-904.

Salzman, N. P. (ed.). 1986. The Papovaviridae: vol. 1: The Polyomaviruses. Plenum Press, New York.

Sambrook, J., Westphal, H., Srinivasan, P. R., and Dulbecco, R. 1968. Proc. Natl. Acad. Sci. USA **60**, 1288-1295.

Sambrook, J., Sugden, B., Keller, W., and Sharp, P. A. 1973. Proc. Natl. Acad. Sci. USA **70**, 3711-3715.

Sambrook, J., Botchan, M., Hu, S.-L., Mitchison, T., and Stringer, J. 1980. Proc. Roy. Soc. **(B) 210**, 423-435.

Sanger, F., Nicklen, S., and Coulson, A. R. 1977. Proc.Natl.Acad.Sci.USA **74**, 5463-5476.

Santangelo, G. M., and Cole, C. N. 1983. Mol. Cell. Biol. **3**, 643-653.

Sasavage, N. L., Smith, M., Gillam, S., Woychick, R. P., and Rottman, F. M. 1982. Proc. Natl. Acad. Sci. USA **79**, 223-227.

Sassone-Corsi, P., Dougherty, J. P., Wasylyk, B., and Chambon, P. 1984. Proc. Natl. Acad. Sci. USA **81**, 308-312.

Sassone-Corsi, P., Wildemann, A., and Chambon, P. 1985. Nature **313**, 458-563.

Sato, K., Ito, R., Baek, K.-H., and Agarwal, K. 1986. Mol. Cell. Biol. **6**, 1032-1043.

Schaffhausen, B. S. 1983. CRC Crit. Rev. Biochem. **13**, 215-228.

Schaffhausen, B. S., and Benjamin, T. L. 1979. Cell **18**, 935-946.

Schaffhausen, B. S., and Benjamin, T. L. 1981. J. Virol. **40**, 184-196.

Scholer, H. R., and Gruss, P. 1984. Cell **36**, 403-411.

Scholer, H. R., and Gruss, P. 1985. EMBO J. **4**, 3005-3013.

Schutzbank, T., Robinson, R., Oren, M., and Levine, A. J. 1982. Cell **30**, 481-490.

Scott, M. R. D., and Simanis, V. 1985. In: DNA Cloning: a Practical Approach. Glover, D. M. (ed.). IRL Press, Oxford, England.

Scott, M. R. D., Westphal, K.-H., and Rigby, P. W. J. 1983. Cell **34**, 557-567.

Sefton, B. M., Hunter, T., Beemon, K., and Eckhart, W. 1980. Cell **20**, 807-816.

Segawa, K., and Ito, Y. 1982. Proc. Natl. Acad. Sci. USA **79**, 6812-6816.

Seif, R., and Martin, R. G. 1979. J. Virol. **32**, 979-988.

Sergeant, A., Bohmann, D., Zentgraf, H., Weiher, H., and Keller, W. 1984. J. Mol. Biol. **180**, 577-600.

Setlow, V. P., Persico-DeLauro, M., Edwards, C. A. F., and Martin, R. G. 1980. Virology **101**, 250-260.

Setzer, D. R., McGrogan, M., Nunberg, J. H., and Schimke, R. T. 1980. Cell **22**, 361-370.

Shelton, E. R., Wassarman, P. M., and DePamphilis, M. L. 1978. J. Mol. Biol. **125**, 491-514.

Shenk, T. E., Carbon, J., and Berg, P. 1976. J. Virol. **18**, 664-672.

Shyamala, M., Atcheson, C. L., and Kasamatsu, H. 1982. J. Virol. **43**, 721-729.

Silver, J., Schaffhausen, B. S., and Benjamin, T. L. 1978. Cell **15**, 485-496.

Simonsen, C. C., and Levinson, A. D. 1983. Mol. Cell. Biol. **3**, 2250-2258.

Smith, A. E., Smith, R., Griffin, B. E., and Fried, M. 1979. Cell **18**, 915-924.

Solnick, D. 1985. Cel **43**, 667-676.

Sompayrac, L., and Danna, K. J. 1983. J. Virol. **46**, 475-480.

Sompayrac, L., and Danna, K. J. 1985. Virology **142**, 432-436.

Spandidos, D. A., and Wilkie, N. M. 1983. EMBO J. **2**, 1193-1199.

Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R., and O'Malley, B. W. 1980. Cell **21**, 681-687.

Stewart, S. E., Eddy, B. E., Gochenour, A. M., Borgese, N. G., and Grubbs, G. E. 1957. Virology **3**, 380-400.

Stoker, M., and Dulbecco, R. 1969. Nature **223**, 397-398.

Streuli, C. H., and Griffin, B. E. 1987. Nature **326**, 619-622.

Stringer, J. R. 1982. Nature **296**, 363-366.

Strub, K., Galli, G., Busslinger, M., and Birnstiel, M. L. 1984. EMBO J. **3**, 2801-2807.

Sweet, B. H., and Hilleman, M. R. 1960. Proc. Soc. Exp. Biol. Med. **105**, 420-427.

Swimmer, C., and Shenk, T. 1984. Proc. Natl. Acad. Sci. USA **81**, 6652-6656.

Takahashi, K., Vigneron, M., Matthes, H., Wildeman, A., Zenke, M., and Chambon, P. 1986. Nature **319**, 121-126.

Takemoto, K. K. 1980. In: Viruses in Naturally Occurring Cancers (Essex, M., Todaro, G., and zur Hausen, H., eds.), pp 311-318. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Tamkun, J. W., Schwarzbauer, J. E., and Hynes, R. O. 1984. Proc. Natl. Acad. Sci. USA **81**, 5140-5144.

Tang, W. J., Berger, S. L., Treizenberg, S. J., and Folk, W. R. 1987. Mol. Cell. Biol. **7**, 1681-1690.

Taya, Y., Devos, R., Tavernier, J., Cheroutre, H., Engler, G., and Fiers, W. 1982. EMBO J. **1**, 953-958.

Tegtmeyer, P. 1972. J. Virol. **10**, 591-602.

Tegtmeyer, P. 1975. J. Virol. **15**, 613-618.

Tegtmeyer, P., Schwartz, M., Collins, J. K., and Rundell, K. 1975. J. Virol. **16**, 168-178.

Templeton, D., Simon, S., and Eckhart, W. 1986. J. Virol. **57**, 367-370.

Tenen, D. G., Taylor, T. S., Haines, L. L., Bradley, M. K., Martin, R. G., and Livingston, D. M. 1983. J. Mol. Biol. **168**, 791-808.

Thompson, J. A., Radonovich, M. F., and Salzman, N. P. 1979. J. Virol. **31**, 437-446.

Tjian, R. 1981. Curr. Top. Microbiol. Immunol. **93**, 5-24.

Tjian, R., and Robbins, A. 1979. Proc. Natl. Acad. Sci. USA **76**, 610-614.

Todaro, G. J., and Green, H. 1964. Virology **23**, 117-119.

Tooze, J. (ed.). 1981. DNA Tumour Viruses (2nd edition), Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Tosi, M., Young, R. A., Hagenbuchle, O., and Schibler, U. 1981. Nucl. Acids Res. **9**, 2313-2323.

Treisman, R. 1980. Nucl. Acids Res. **8**, 4867-4888.

Treisman, R. 1981. PhD Thesis. University of London.

Treisman, R. H., Cowie, A., Favaloro, J. M., Jat, P., and Kamen, R. 1981a. J. Mol. Appl. Genet. **1**, 83-92.

Treisman, R., Novak, U., Favaloro, J., and Kamen, R. 1981b. Nature **292**, 595-600.

Treisman, R., and Maniatis, T. 1985. Nature **315**, 72-75.

Turler, H., and Salomon, C. 1985. J. Virol. **53**, 579-586.

Twigg, A. J., and Sherratt, D. 1980. Nature **283**, 216-218.

Van der Eb, A. J., Van Ormondt, H., Schrier, P. I., Lupker, J. H., Jochemsen, H., Van der Elsen, P. J., Deleys, R. J., Maat, J., Van Deveren, C. P., Dijkema, R., and De Waard, A. 1979. Cold Spring Harbor Symp. Quant. Biol. **44**, 383-399.

Varmus, H. E. 1984. Ann. Rev. Genet. **18**, 533-612.

Varmus, H. E., and Swanstrom, R. 1982. In: RNA Tumor Viruses (2nd. Ed.). 369-512. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Vigneron, M., Barrerra-Saldana, H. A., Baty, D., Everett, R. E., and Chambon, P. 1984. EMBO J. **3**, 2373-2382.

Vogt, M., and Dulbecco, R. 1960. Proc. Natl. Acad. Sci. USA **46**, 365-370.

Walter, G., Hutchinson, M. A., Hunter, T., and Eckhart, W. 1982. Proc. Natl. Acad. Sci. USA **79**, 4025-4029.

Ward, D. F., and Murray, N. E. 1979. J. Mol. Biol. **133**, 249-266.

Wasylyk, B., Wasylyk, C., Matthes, H., Wintzerith, M., and Chambon, P. 1983. EMBO J. **2**, 1605-1611.

Weber, F., de Villiers, J., and Schaffner, W. 1984. Cell **36**, 983-992.

Weiderman, L. M., and Perry, R. P. 1984. Mol. Cell. Biol. **4**, 2518-2528.

Whitelaw, E., and Proudfoot, N. J. 1986. EMBO J. **5**, 2915-2922.

Wickens, M., and Stephenson, P. 1984. Science **226**, 1045-1051.

Wildeman, A. G., Sassone-Corsi, P., Gundstrom, T., Zenke, M., and Chambon, P. 1984. EMBO J. **3**, 3129-3133.

Williams, J. G., Hoffman, R., and Penman, S. 1977. Cell **11**, 901-907.

Williams, T., and Fried, M. 1986. Mol. Cell. Biol. **6**, 2179-2184.

Wilson, J.B. 1986. Ph.D. Thesis, University of London.

Wilson, J. B., Hayday, A., Courtneige, S., and Fried, M. 1986. Cell **44**, 477-487.

Winston, V. D., Bolen, J. B., and Consigli, R. A. 1980. J. Virol. **33**, 1173-1181.

Wold, M. S., Li, J. J., and Kelly, T. J. 1987. Proc. Natl. Acad. Sci. USA **84**, 3643-3647.

Woychick, R. P., Lyons, R. H., Post, L., and Rottman, F. M. 1984. Proc. Natl. Acad. Sci. USA **81**, 3944-3948.

Yanofsky, C. 1981. Nature **289**, 751-758.

Yuan, D., and Tucker, P. W. 1984. J. Exp. Med. **160**, 564-583.

Yuen, L. K. C., and Consigli, R. A. 1985. Virology **144**, 127-138.

Zaret, K. S., and Sherman, F. 1982. Cell **28**, 563-573.

Zaret, K. S., and Sherman, F. 1984. J. Mol. Biol. **176**, 107-135.

Zarkower, D., Stephenson, P., Sheets, M., and Wickens, M. 1986. Mol. Cell. Biol. **6**, 2317-2323.

Zarkower, D., and Wickens, M. 1987. EMBO J. **6**, 177-186.

Zenke, M., Gundstrom, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A., and Chambon, P. 1986. EMBO J. **5**, 387-397.

Zhu, Z., Veldman, G. M., Cowie, A., Carr, A, Schaffhausen, B., and Kamen, R. 1984. J. Virol. **51**, 170-180.

Zuckerman, S.H., Linder, S., and Eisenstadt, J.M. 1984. Biochim. Biophys. Acta **804**, 285-290.

APPENDIX 1: GENE ACTIVATION PROPERTIES OF A MOUSE DNA SEQUENCE ISOLATED BY EXPRESSION SELECTION

Volker von Hoyningen-Huene, Chris Norbury, Moira Griffiths and Mike Fried.

Published 1986. Nucl. Acids Res. **14**, 5615-5627.

Volume 14 Number 14 1986 Nucleic Acids Research

Gene activation properties of a mouse DNA sequence isolated by expression selection

Volker von Hoyningen-Huene, Chris Norbury, Moira Griffiths and Mike Fried

Department of Tumour Virus Genetics, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK

Received 29 May 1986; Accepted 2 July 1986

ABSTRACT

The MES-1 element was previously isolated from restricted total mouse cellular DNA by "expression selection" - the ability to reactivate expression of a test gene devoid of its 5' enhancer sequences (1). Mes-1 has been tested in long-term transformation and short-term CAT expression assays. In both assays MES-1 is active independent of orientation and at a distance when placed 5' to the test gene. The element is active with heterologous promoters and functions efficiently in both rat and mouse cells. MES-1 activates expression by increasing transcription from the test gene's own start (cap) site. Thus the expression selection technique can be used for the isolation of DNA sequences with enhancer-like properties from total cellular DNA.

INTRODUCTION

Transcriptional enhancers are small sequence elements which can function from a distance in an orientation-independent fashion to increase the expression of a gene via initiation of transcription from the gene's own start (cap) site. Although enhancer elements were initially identified in viruses (2-4), more recently a number of sequence elements with some of the properties of enhancers have been found associated with cloned cellular genes (for reviews see 5,6). We have taken an alternative approach called "expression selection" for the isolation of cellular enhancer sequences (1,7). This relies on the ability of restriction fragments of genomic DNA to confer expression upon a selectable gene devoid of its own enhancer element. A 149 bp fragment of mouse cell DNA formerly called H-2 and henceforth referred to as MES-1 (mouse expression sequence 1) was isolated using this expression selection technique (1). In this report the ability of MES-1 to enhance the expression of heterologous genes lacking an enhancer element has been characterized using long-term and short-term expression assays. MES-1 has been found to share a number of properties with classical enhancer elements.

MATERIALS AND METHODS

Plasmids

All plasmids contain the ampicillin resistance gene of pBR322 and were propagated in E.coli strain DH-1 (8). The supercoiled plasmids used for DNA transfection were purified twice by centrifugation, first by density in CsCl and second by velocity sedimentation through sucrose.

Expression Assays

RAT-1 and Ltk(-) cells were used for short-term CAT assays. Cells were plated the day before transfection onto 5 cm petri dishes containing 2×10^5 cells in Eagles medium containing 10% fetal calf serum. The medium was changed 4 hours prior to transfection. DNA transfections and assay of CAT activity were carried out essentially as described by Gorman et al. (9). Plasmid DNA was precipitated at a concentration of 20 µg/ml in the transfection mix. Ten µg of plasmid DNA were transfected onto each petri dish containing a total volume of 5 ml medium. Cells were incubated with the precipitate for 12 hours at 32°C, washed twice with versene and provided with fresh medium. Cells were then incubated for 48 hours at 37°C after which time the cells were harvested and assayed for CAT activity. The results of CAT reactions were visualised by autoradiography of the TLC plates and the reaction products thus localized were extracted and quantified by scintillation counting (9). The long-term transformation assays using Rat-1 cells were performed as described by Fried et al. (1). The cells were stained with Leishman's stain seventeen days after transfection and the number of densely packed transformed foci counted.

Transient Expression/S1 Nuclease Protection Assays

Ten ml of calcium phosphate precipitate containing 15 µg/ml supercoiled plasmid was transfected onto 5×10^6 Rat-1 cells on a bioassay tray containing 90 ml Eagles medium supplemented with 10% fetal calf serum. After incubation at 37°C for 5 hours, the precipitate was removed by washing with versene and the cells were refed with fresh medium. Forty hours later total cellular RNA was prepared by lysis with guanidinium isothiocyanate as described (8). An AccI digest of pPy70 was dephosphorylated using calf intestinal alkaline phosphatase (Boehringer) and the 5' ends labeled using γ-^{32}P ATP (3000 Ci/mmol, Amersham) and polynucleotide kinase (Boehringer). After digestion with BamHI, the 401 bp double stranded BamHI-AccI fragment spanning Py nts. 5265-371 was isolated by agarose gel electrophoresis and used as a probe. Forty µg of total cellular RNA

was annealed with 40 000 dpm of probe in 20 µl of hybridisation buffer as described (10). S1 digestion was performed for 2 hours at 14°C with 500 units/ml S1 nuclease (Sigma) and resistant fragments were separated on a 5% polyacrylamide/7M urea sequencing-type gel.

RESULTS

Vectors Used for Expression Assays

The 149 bp MES-1 sequence was originally isolated as a BamHI-HaeII restriction fragment (1). The HaeII cleavage site was changed to a BglII restriction site by linker addition in order to facilitate easy insertion, in both orientations, of the fragment into the vectors used for the long- and short-term expression assays. These vectors are shown in Fig. 1.

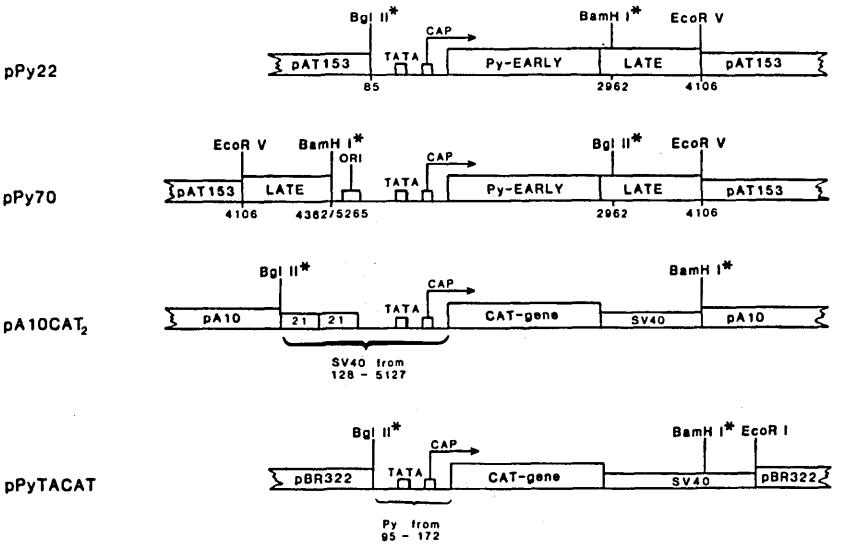


Fig. 1. Expression vectors. The relevant regions of the plasmids pPy22 and pPy70 (long-term transformation vectors) and pA10CAT2 and pPyTACAT (short-term CAT expression vectors) described in the text are shown. The BamHI and BglII restriction sites used for insertion of the MES-1 element are indicated by asterisks (*) and the direction of transcription of the respective test genes by arrows. The plasmid parts of the respective vectors (pAT153, pA10, pBR322), the Py origin of DNA replication (ORI), the Py or SV40 TATA box sequences (TATA) and cap sites (CAP) and the other Py and CAT sequences are indicated. The parts of the CAT vectors labelled SV40 3' to the CAT gene contain the SV40 early region splice and poly-adenylation signal. Numbering under the pPy22 and pPy70 vectors refers to Py nucleotide numbers.

The pPy70 vector used for the long-term transformation assays contains polyoma virus (Py) DNA linearised at the EcoRV site and lacking 633 bp of Py sequence between the BamHI site at nt 4632 and the PvuII site at nt 5265. The PvuII site at nt 5265 has been converted to a BamHI site by linker addition. The pPy22 long-term transformation vector contains 4.0 kb of Py sequence, between the NarI site at Py nt 85 and the EcoRV site at Py nt 4106, which has been cloned into the pAT153 plasmid between the EcoRV site at nt 185 and the NarI site at nt 1204. The single NarI site in pPy22 has been converted to a BglII site by linker addition. The Py sequences in both vectors contain the complete Py early region (transforming region) with its associated TATA box and cap site and part of the Py late region (11), (pPy70 also contains the origin of viral DNA replication), but lack the Py enhancer region (nts. 5021-5265) (12) and are thus greatly inhibited in their transforming ability (1,7). For insertion of the BamHI-BglII MES-1 fragment 5' to the Py transforming gene the BamHI site at Py nt 5265 in pPy70 or the BglII site at Py nt 85 in pPy22 was used. These sites are about 65 bp (pPy22) and 180 bp (pPy70) upstream of the major sites used for the initiation of transcription of the Py early mRNAs (cap sites) (13). For insertion of the MES-1 fragment 3' to the transforming region the HincII site at Py nt 2962, just downstream of the Py early region, was used after conversion to a BglII site in pPy70 or to a BamHI site in pPy22 by linker addition.

The plasmids pA10CAT2 (14) and pPyTACAT were used for the short-term assays (Fig. 1). The pA10CAT2 plasmid contains the bacterial chloramphenicol acetyl transferase (CAT) gene which is preceded by the SV40 promoter element containing the 21 bp repeats and followed by the SV40 early region splice and polyadenylation signals. The enhancerless pA10CAT2 plasmid is greatly inhibited in inducing the production of CAT activity after transfection into mammalian cells (14). The BamHI-BglII MES-1 fragment was inserted 5' to the CAT gene at the BglII cleavage site, at the join of the SV40 promoter element with the plasmid sequences, which is about 140 bp upstream from the major cap sites of the SV40 early region. In the pPyTACAT vector a Py fragment containing the Py early region TATA box and cap sites has been inserted 5' to the CAT gene of the pSVOCAT plasmid of Gorman et al. (9). The HaeII site at Py nt 95 and the BstXI site at Py nt 166 at the ends of this Py fragment were converted into BglII and BamHI sites respectively by linker addition and inserted into the BglII site (converted from the original HindIII site) of pSVOCAT so

Table 1 Effect of the MES-1 fragment on expression of long-term transformation vectors

	Plasmid	transformed foci per 100 ng of DNA
	pPy22	3
→	pPy22	184
←	pPy22	37
	pPy22 →	5
	pPy22 ←	4
	pPy70	3
→	pPy70	172
←	pPy70	191
	pPy70 →	3
	pPy70 ←	7
	pPy	192

The arrows indicate the orientation of the MES-1 element when placed either 5' (before) or 3' (after) in either the pPy70 or pPy22 vectors. The arrow pointing from left to right represents the orientation in which MES-1 was originally isolated. pPy is a plasmid that contains the Py early region and its associated enhancer element.

that the BglII site in front of the Py fragment was maintained (Fig. 1). The MES-1 fragment was inserted 5' to the CAT gene at this BglII site which is about 55 bp upstream from the Py early region major cap sites. In both pA10CAT2 and pPyTACAT the MES-1 fragment was inserted 3' to the CAT gene at the BamHI site just 3' to the SV40 polyadenylation signal (Fig. 1).

MES-1 Expression Activity

The results of the induction of transforming activity by MES-1 when placed in either orientation either 5' or 3' to the Py transforming region in pPy70 and pPy22 are presented in Table 1. When placed in the 5' position in pPy70, MES-1 induces an orientation-independent increase in transforming activity of approximately 60 fold. No increase in activity

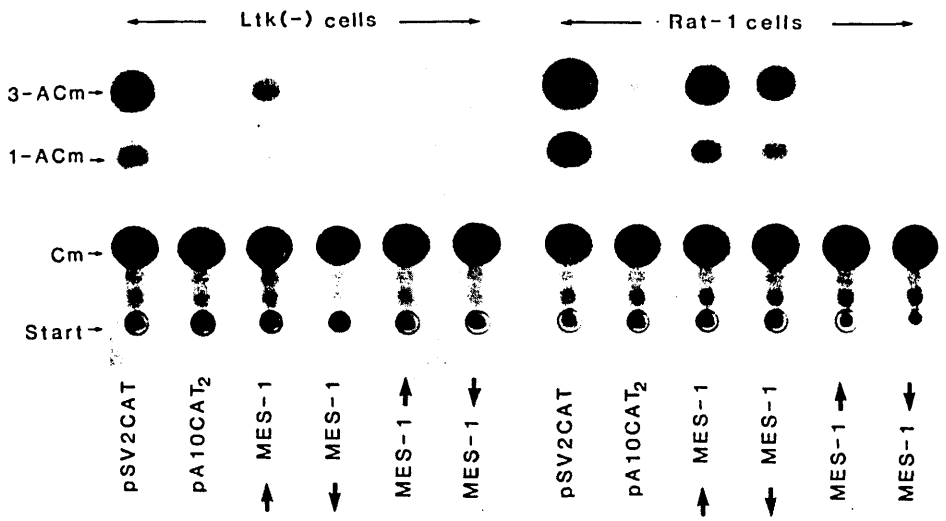


Fig. 2. CAT assay using the pA10CAT2 vector. The MES-1 sequence was inserted into pA10CAT2 either at the 5' or 3' position to the CAT gene (see Fig 1) and assayed after transfection into mouse Ltk(-) or Rat-1 cells. The pSV2CAT plasmid containing the SV40 enhancer (9) and the enhancerless pA10Cat2 (14) served as positive and negative controls respectively. An autoradiogram of the reaction products is shown after fractionation on a TLC plate. The orientation of MES-1 is indicated by the arrows with the arrow going from left to right indicating the orientation of MES-1 as originally isolated (1). Cm = Chloramphenicol, 1-ACm, 3-ACm = 1- or 3- Acetylchloramphenicol.

was detected when MES-1 was inserted in pPy70 in either orientation at a single position 3' to the transforming region. The MES-1 fragment also increased transforming activity in an orientation-independent manner when placed in a 5' position in the pPy22 vector, but the increase in activity in one orientation was always 3-10 fold greater than that in the other orientation.

The MES-1 fragment also induced an increase in CAT activity in the short-term expression assays in both rat and mouse cells (Fig. 2 and Table 2). Whereas there was only a 2-5 fold difference in increased activity between the two orientations with pA10CAT2, the increase in activity in pPyTACAT was 7 to 50 fold higher in one orientation than in the other (Table 2). The orientation showing the greater activity in pPyTACAT was the same orientation that showed an increased activity with pPy22 in the long-term expression assays (Table 1). Little or no increase in CAT act-

Table 2 Enhancement of CAT gene expression by the MES-1 sequence

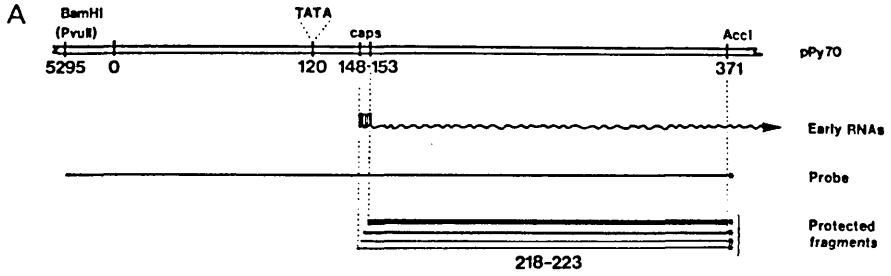
Plasmid	relative enhancement in Rat-1 cells	Ltk(-) cells
pA10CAT2	1	1
→ pA10CAT2	12	22
← pA10CAT2	9	9
pA10CAT2 →	1	1
pA10CAT2 ←	1	1
pPyTACAT	1	1
→ pPyTACAT	33	98
→ pPyTACAT	5	2
pPyTACAT →	1	1
pPyTACAT →	1	1
pSV2CAT	45	43

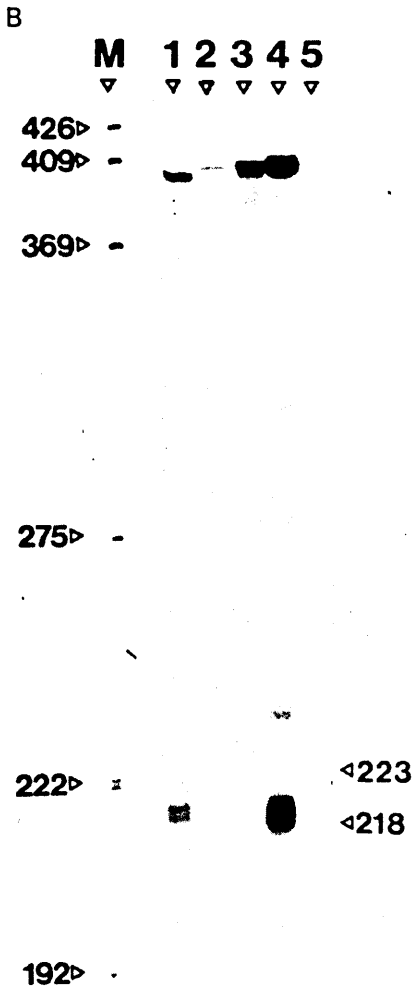
The arrows indicate the orientation of the MES-1 element when placed either 5' (before) or 3' (after) in either the pA10CAT2 (Laimins et al. 1984) or pPyTACAT vectors. The arrow pointing from left to right represents the orientation in which MES-1 was originally isolated. Relative enhancement is shown with respect to the vectors alone. pSV2CAT (Gorman et al. 1982) contains the early region promoter and enhancer of SV40 in front of the CAT gene.

ivity was detected when MES-1 was inserted in either orientation at a single 3' position in the two vectors (Fig. 2 and Table 2).

MES-1 Induces Initiation of Transcription at the Py Cap Site

In the cell line, isolated after "expression selection," from which the MES-1 element was molecularly cloned, reactivation of transcription of the selectable test gene was shown to be initiated at the correct start (cap) sites (1). This could have been due either to the MES-1 element or another sequence in the vicinity of the integrated test gene. The finding that the isolated MES-1 element induced an increase in biological


A
BamHI
(PvuII)
TATA
caps
AccI
pPy70
5295
0
120
148-153
371
Early RNAs
Probe
Protected
fragments
218-223


B
M
1 2 3 4 5
426
409
369
275
222
223
218
192

activity of a test gene, in an orientation-independent manner, when positioned at different distances 5' to the test gene (see above) indicated that MES-1 contains gene enhancer activity. Thus it would be expected that the MES-1-induced increase in biological activity would be due to an increase in transcription and that this increased transcription would be initiated at the test gene's cap site.

The positions and relative abundance of the 5' ends of transcripts induced by MES-1 were determined after transient expression in rat cells. Transcripts obtained by insertion of MES-1 into the pPy70 vector in both orientations are compared with those from the vector alone and from a similar plasmid containing the Py early region and Py enhancer (Fig. 3). The major cap sites for transcription of the Py early region show microheterogeneity and have been mapped between Py nts 148-153 (13). In both orientations MES-1 induced a substantial increase in transcription initiated from the bona fide Py cap sites (Py nts 148-153) compared to the vector alone. A larger amount of transcription was consistently observed in the transient assay when MES-1 was in one orientation (opposite to that as originally isolated) in the pPy70 vector (Fig. 3). In the long term transformation assay no significant differences were observed between the two orientations of MES-1 in the pPy70 vector (Table 1). This apparent discrepancy may reflect the different natures of the two assays. The transient assay measures the average transcriptional activity of a substantial number of cells soon after transfection. In contrast in the long-term assay a stably transformed phenotype is monitored in a small fraction of the cells weeks after transfection. Attempts to compare the two assays may consequently lead to artificial contradictions.

Fig. 3. MES-1-induced transcription from the Py early cap sites. (A) The transcription initiation region of the pPy70 vector aligned with the major early capped Py RNAs, initiated at nts.148-153 (13). The 401 bp fragment between Py nts. 5265 and 371 was used as a probe in an S1 nuclease mapping experiment after 5' end labelling at nt.371. Py RNAs using the bona fide Py cap sites protect fragments in the range 218-223 nucleotides. (B) S1-resistant fragments produced by annealing the probe with RNAs transiently expressed from (1) a plasmid pAT153 clone containing a BamHI-linearised Py molecule (contains Py enhancer). (2) The pPy70 vector lacking the Py enhancer. (3) pPy70 containing the 149 bp MES-1 sequence inserted at the pPy70 BamHI site with the MES-1 fragment in the orientation as originally isolated (1). (4) pPy70 containing the 149 bp sequence in the opposite orientation. (5) The probe annealed to yeast tRNA. Size markers (M) are end-labelled DdeI fragments of the pAT153 plasmid containing the BamHI-Linearised Py molecule.

DISCUSSION

The isolation of the 149 bp MES-1 element from mouse genomic DNA was facilitated by the element's ability to reactivate a selectable gene devoid of its 5' expression sequences (1). The results presented here show that MES-1 has many of the properties associated with gene enhancer elements. MES-1 can activate a test gene in an orientation-independent manner when placed at least 180 bp upstream from the cap site of the test gene. The MES-1-induced increase in biological activity is the result of an increase in transcription and the MES-1-induced transcripts are primarily initiated at the test gene's own cap site. MES-1 can activate heterologous (Py and SV40) "promoters" in both rat and mouse cells. In addition MES-1 has also been shown to be active in dog, monkey and human cells (V. von Hoyningen-Huene unpublished results). The MES-1 element does not appear to contain any of the consensus sequences associated with other classes of enhancer elements (1) nor does it appear to contain the consensus sequence for the Sp1 factor binding sites (15).

The activity of MES-1 appears to be sensitive to sequences in the vicinity of its location. Whereas MES-1 was not very active in one orientation when placed 55 bp (pPyTACAT) or 65 bp (pPy22) from the Py cap site it was highly active in the same orientation when placed 180 bp (pPy70) from the Py cap site or 140 bp in front of the SV40 cap site in pA10CAt (Fig. 2, Tables 1 and 2). This sensitivity to location may explain why MES-1 has not been found to be active in the two positions tested at a 3' position to a test gene. Alternatively, MES-1 may represent a class of cellular enhancer-like sequences which have only some of the classical gene enhancer properties defined initially in viruses. The activities of enhancer-like sequences associated with the human c-fos gene (16), the rat insulin I gene (17) and the herpes simplex virus intermediate early gene 3 (18, 19) have also been found to be sensitive to position and orientation. For instance the 5' regulatory sequences of the c-fos gene are active in an orientation-independent manner only at a limited distance 5' to the c-fos gene and have little or no activity when placed 3' to the c-fos gene (16).

Recent results indicate that a given sequence can have both enhancer and promoter activities (20, 21). The promoter sequences associated with the Py early region have not yet been defined. Thus it is possible that in at least some of the Py constructs MES-1 is displaying promoter-like activity. Although some promoter elements can be active independent of

orientation they do not appear to be active when moved away from the test gene (22,23). The MES-1 element is active both close to and at a distance from the test gene in the Py constructs (Tables 1 and 2). In the case of the pA10CAT2 plasmid which contains the SV40 21 bp repeat promoter element MES-1 induces an orientation-independent increase in CAT activity, when placed in front of the SV40 promoter (Fig. 2 and Table 2), as would be expected for an enhancer sequence. The enhancer/promoter activity of MES-1 in its natural chromosomal location in the mouse genome awaits further elucidation.

The "expression selection" technique was developed in order to be able to isolate expression sequences independent of their association with a cloned gene (1,7). The "enhancer trap" method of Weber et al. (24) was also devised in order to select and isolate enhancer elements. The "enhancer trap" method relies on intracellular recombination between transfected enhancer sequences and an enhancerless SV40 vector to produce an infectious SV40 molecule. There is a constraint on the size of the DNA incorporated in the vector as the hybrid molecule must be packaged into SV40 virions. The incorporated enhancer must be able precisely to regulate the SV40 early and late region transcription so that the hybrid viral DNA can efficiently replicate and be packaged into new infectious virus particles during the lytic cycle in monkey cells. Using this method the enhancer sequences once covalently linked to the SV40 DNA are easily isolated from the abundant replicated extrachromosomal circular SV40 DNA molecules (24).

In the "expression selection" method the enhancer sequence and the test gene are recombined (ligated) in vitro. The hybrid DNA molecules are then transfected into cells and those expressing the test gene from a chromosomal (integrated) location are selected and analysed. There are no size constraints on the DNA that contains the enhancer element. All that is required of the enhancer element is that it restores activity to the test gene in its chromosomal location. In contrast the "enhancer trap" requires that a number of temporal events be precisely regulated by the enhancer to produce infectious SV40 virus. The isolation of enhancer like sequences detected by "expression selection" requires laborious cloning from cellular DNA (1,7) as opposed to the easy isolation of such sequences from abundant extrachromosomal circular DNA (see above) using the "enhancer trap" method. Although the "enhancer trap" method has proven very useful in detecting enhancer elements when large amounts of

a cloned DNA, especially from viruses, are transfected with the enhancerless SV40 vector (20, 24-28), this method has not been successful in selecting enhancer elements from total cellular DNA (20, 24). In contrast the "expression selection" method has been successful in selecting sequences with enhancer properties from total cellular DNA (1,7) possibly as a result of the less stringent requirements for gene activation.

Once isolated an enhancer-like sequence can be used as a probe for the cloning of surrounding DNA sequences which may contain associated "companion" genes. The finding that MES-1 is located between the 5' ends of two divergent transcription units (T. Williams and M. Fried manuscript submitted) is consistent with the enhancer-like activity of the MES-1 sequence presented here. Furthermore a second sequence isolated by "expression selection" by Ford et al. (7) also appears to contain a gene enhancer sequence. Thus the use of "expression selection" appears to be a fruitful technique for the isolation of cellular expression sequences with enhancer activity.

ACKNOWLEDGEMENTS

We are grateful to Trevor Williams and Nicolas C. Jones for their critical comment and to Georgina Briody for typing the manuscript.

REFERENCES

1. Fried, M., Griffiths, M., Davies, B., Bjursell, G., LaMantia, G. and Lania, L. (1983) Proc. Natl. Acad. Sci. USA **80**, 2117-2121.
2. Banerji, J., Rusconi, S. and Schaffner, W. (1981) Cell. **27**, 299-308.
3. Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gau, M. P. and Chambon, P. (1981) Nucleic Acids Res. **9**, 6047-6068.
4. Fromm, M. and Berg, P. (1983) Mol. Cell. Biol. **3**, 991-999.
5. Gluzman, Y., ed. (1985) Eukaryotic Transcription: The Role of Cis- and Transacting Elements in Initiation. Published by Cold Spring Harbor Laboratory Press, NY.
6. Serfling, E., Jasin, M. and Schaffner, W. (1985) Trends Genet., **1**, 224-230.
7. Ford, M., Davies, B., Griffiths, M., Wilson, J. and Fried, M. (1985) Proc. Natl. Acad. Sci. USA **82**, 3370-3374.
8. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Manual. Published by Cold Spring Harbor.
9. Gorman, C., Moffat, L. and Howard, B. (1982) Mol. Cell. Biol. **2**, 1044-1051.
10. Weaver, R. F., and Weissmann, C. (1979) Nucleic Acids Res., **7**, 1175-1193.
11. Tooze, J., ed. (1981) DNA Tumor Viruses (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
12. De Villiers, J. and Schaffner, W. (1981) Nucleic Acids Res. **9**, 6251-6264.
13. Cowie, A., Jat, P. and Kamen, R. (1982) J. Mol. Biol. **159**, 225-255.
14. Laimins, L. A., Gruss, P., Pozatti, R. and Khoury, G. (1984) J. Virol. **49**, 183-189.

15. Dynan, W. S. and Tjian, R. (1985) Nature, **316**, 774-778.
16. Treisman, R. (1985) Cell, **42**, 889-902.
17. Edlund, T., Walker, M. D., Barr, P. J., and Rutter, W. J. (1985) Science, **230**, 912-916.
18. Lang, J. C., Spandidos, D. A. and Wilkie, N. M. (1984) EMBO Journal **3**, 389-395.
19. Preston, C. M. and Tannahill, D. (1984) Virology **137**, 439-444.
20. Serfling, E., Lubbe, A., Dorsch-Hasler, K. and Schaffner, W. (1985) EMBO Journal **4**, 3851-3859.
21. Goodbourn, S., Zinn, K. and Maniatis, T. (1985) Cell **41**, 509-520.
22. Everett, R. D., Baty, D. and Chambon, P. (1983) Nucleic Acids Res. **11**, 2447-2464.
23. McKnight, S. L., Kingsbury, R. C., Spence, A. and Smith, M. (1984) Cell, **37**, 253-262.
24. Weber, F., de Villiers, J. and Schaffner, W. (1984) Cell, **36**, 983-992.
25. Boshart, M., Weber, F., Jahn, G., Dorsch-Hasler, K., Fleckenstein, B. and Schaffner, W. (1985) Cell **41**, 521-530.
26. Dorsch-Hasler, K., Keil, G. M., Weber, F., Jasin, M., Schaffner, W. and Koszinowski, U. H. (1985) Proc. Natl. Acad. Sci. USA **82**, 8325-8329.
27. Schirm, S., Weber, F., Schaffner, W. and Fleckenstein, B. (1985) EMBO Journal **4**, 2669-2674.
28. Weber, F. and Schaffner, W. (1984) EMBO Journal **4**, 949-956.

APPENDIX 2: THE ORGANIZATION AND EXPRESSION OF INTEGRATED POLYOMAVIRUS DNA IN TRANSFORMED CELLS

M. Fried, J.B. Wilson, T. Williams and C. Norbury.

Published 1986. Cancer Cells **4**, 309-316.

The Organization and Expression of Integrated Polyomavirus DNA in Transformed Cells

M. Fried, J.B. Wilson, T. Williams, and C. Norbury

Department of Tumour Virus Genetics, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, United Kingdom

Polyomavirus DNA is found integrated into the cellular chromosome in transformed cells. A number of properties of polyoma integration are described, including the illegitimate recombination events and gross and local rearrangements of cellular DNA sequences that occur at polyoma integration sites. The transformed phenotype is mediated by the polyoma oncogene—middle T antigen (TAg). Three unstable, flat-cell revertants of a polyoma-transformed cell line have been found to contain a deletion of one of nine consecutive cytidine residues between polyoma nucleotides 1239 and 1247 in a region of the polyoma early region where there are two translated reading frames, one coding for the carboxyl terminus of middle TAg and the other for the central region of large TAg. This frameshift mutation results in the shuffling of the different domains of the polyoma TAg's. The 43K protein generated by this shuffling contains the aminoterminal 75% of middle TAg, has an associated protein kinase activity, and forms a complex with c-*src* but is unable to induce a transformed phenotype. Transient frameshifting during normal polyoma gene expression as a mechanism to give the virus greater genetic flexibility is discussed.

Polyomavirus can transform cells to a neoplastic state in vitro and induce tumors in susceptible animals. The transforming ability is mediated by the viral early region, which encodes the three (large, middle, and small) tumor antigens (TAg's). The three TAg's are translated from three different mRNAs, which, although transcribed from the same early region DNA sequence, vary as a result of differential splicing (Tooze 1981) (Fig. 1). The large TAg, which has a molecular weight of 100,000 (100K) and is predominantly located in the nucleus, is a phosphoprotein with DNA-binding activity. The large TAg is required for the initiation of viral DNA replication (Fried 1970; Francke and Eckhart 1973) and can negatively regulate transcription of the polyoma early region mRNAs (Cogen 1978). It is also capable of "immortalizing" primary cells (Rassoulzadegan et al. 1983), and its activity is required for the initiation, but not the maintenance, of the transformed phenotype after virus infection (Fried 1965). The 56K middle TAg, which is the polyoma oncogene (Treisman et al. 1981), has a hydrophobic carboxyl terminus and is found associated with membranes (Ito et al. 1977; Silver et al. 1978). Middle TAg has an associated protein kinase activity (Eckhart et al. 1979; Schaffhausen and Benjamin 1979; Smith et al. 1979), which may be the result of the kinase activity of the c-*src* protein with which it can form a complex (Courtneidge and Smith 1983, 1984). Little is known of the biological activity of the 22K small TAg, which appears to be located in the cytoplasm. In polyoma-transformed cells the viral DNA is invariably found integrated into the host-cell chromosome. In this report we present some of the properties of the organization and expression of the integrated polyoma sequences in transformed cells.

Results and Discussion

Properties of polyomavirus integration

A great deal of information concerning polyomavirus integration has been generated in our laboratory from studies of the integrated polyoma and adjacent cellular DNA sequences of the five polyoma-transformed cell lines shown in Figure 2. These five cell lines (82-Rat, 53-Rat, 7axT, 7axB, and TSA-3T3) all contain a single insert of polyoma DNA, and in some cases the viral insert and adjacent cellular DNA sequences have been molecularly cloned and analyzed (Hayday et al. 1982; Ruley et al. 1982, 1986; Ruley and Fried 1983). A number of the properties of polyoma integration are presented in Table 1. From an inspection of virus-host joins (Figs. 2 and 3), it can be seen that integration can occur at many different locations in the host-cell DNA or viral DNA.

In contrast to retroviruses such as avian leukemia viruses and mouse mammary tumor virus, the transformed phenotype induced by polyomavirus does not appear to be caused by integration in specific regions of the cellular genome and consequent activation of an adjacent cellular gene (for review, see Varmus 1984). Neither does polyoma transformation appear to occur as the result of the inactivation of a cellular gene, because an insertion in only one cellular chromosomal allele (haploid insertion) is sufficient for expression of the transformed phenotype (Hayday et al. 1982). Polyoma

CANCER CELLS 4 / *DNA Tumor Viruses.* 0-87969-192-1/86 $1.00

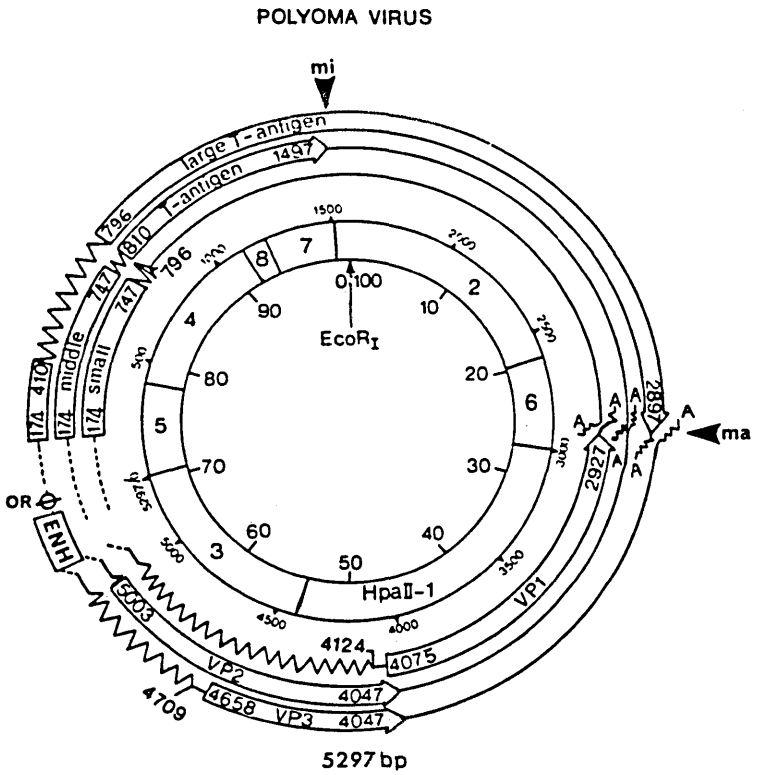


Figure 1 The polyomavirus map of the A-2 strain (5297 bp). The early and late region mRNAs (thin lines) and their coding regions (boxed regions) are shown relative to the *HpaII* physical map. Map units are shown inside and nucleotide numbers outside the physical map. The nucleotide numbers at the starts and ends of the coding regions and at the splice donors and acceptors are shown. Introns are indicated by jagged lines. The positions of the origin of DNA replication (OR), the enhancer region (ENH), and the major (*ma*) and minor (mi) early region polyadenylation sites are shown. Modified from Tooze (1981).

transformation seems, rather, to be the result of the addition of virally specified genetic information to the cellular genome. In transformed cells the polyoma early region is always expressed whereas the polyoma late region is poorly expressed, if at all. Although after virus infection a functional polyoma large TAg is required for the initiation of transformation (Fried 1965), upon continued growth of the transformed cells both in vitro (Basilico et al. 1979, 1980) and in vivo (Lania et al. 1981) there appears to be a selection for a nonfunctional large TAg by either deletion or mutation of the early region DNA sequences that are unique to the coding region of large TAg (polyoma nucleotides [nt] 1498–2897; see Fig. 1). These sequences are truncated in the polyoma-transformed cell lines 82-Rat, 53-Rat, 7axT, 7axB, and TSA-3T3 as a result of recombination between cellular and viral sequences (Fig. 2). The intact early regions of 7axB contain a single-base-change mutation in the sequences unique to large TAg that inactivates large TAg function (Hayday et al. 1983), whereas one of the early regions of 82-Rat contains a deletion-insertion in the sequences unique to large TAg (Ruley and Fried 1983; Ruley et al. 1986). Cells containing integrated polyoma host-range transformation (*hr-t*) mutant DNA, capable of synthesizing a functional large TAg but not a functional middle TAg, have a normal phenotype (Lania et al. 1979), indicating that integration per se or the expression of a functional large TAg is not sufficient for the induction of a transformed phenotype. In contrast, the transformed phenotype requires an intact, functional middle TAg coding region. An intact middle TAg coding region is always present and is always expressed in polyoma-transformed cells (Lania et al. 1980) (Fig. 2), and a middle TAg cDNA will efficiently transform established cells (Treisman et al. 1981). In a number of cell lines, the recombinant join has resulted in the loss of the 3′ end of the polyoma early region and its associated polyadenylation signal. This can result in transcripts initiated at the start of the viral early region continuing into the adjacent host DNA where their polyadenylation is dependent on a cellular processing signal (Ruley et al. 1982).

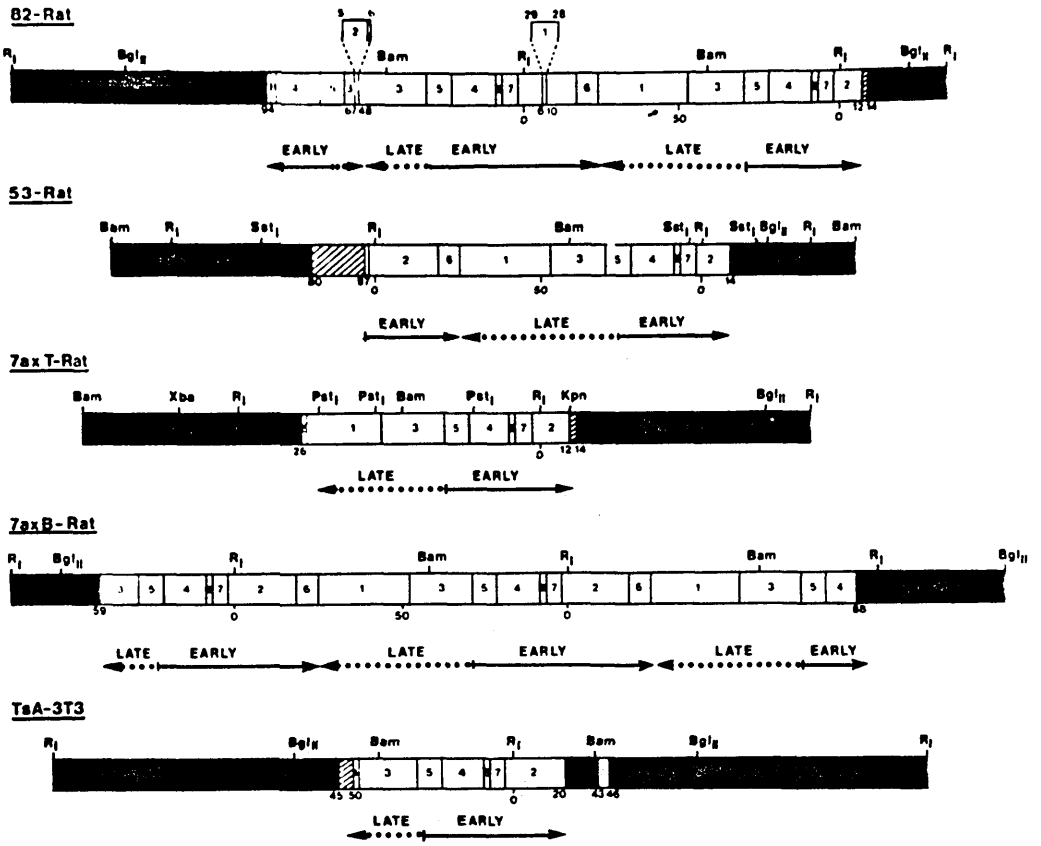


Figure 2 The physical maps of the viral and adjacent cellular DNA of five polyoma-transformed cell lines. The *Hpa*II maps of the integrated viral sequences are shown in white (see Fig. 1), and the flanking cellular sequences are in black for the cell lines 82-Rat (Ruley and Fried 1983; Ruley et al. 1986), 53-Rat (Ruley et al. 1986), 7axT (Hayday et al. 1982), 7axB (Hayday et al. 1982), and TSA-3T3 (Ruley et al. 1982). Regions of uncertainty at the virus-host joins are indicated by hatching. The viral early and late regions are shown.

Table 1 Polyomavirus Integration

1. Integration appears to take place in many different regions of the host cell and viral DNA. There is no evidence of integration in specific regions of cellular DNA.
2. A haploid insertion of viral DNA is sufficient for the transformed phenotype.
3. Cellular DNA adjacent to the viral insert does not contain transforming activity; viral transformation does not appear to activate cellular oncogenes.
4. The transformed phenotype requires an intact, functional middle TAg coding region. An intact large TAg coding region is not essential for the maintenance of the transformed phenotype after viral integration, although it is required for initiation of transformation by the infecting virus.
5. There appears to be selection for loss of functional large TAg both in vitro and in vivo either by mutation or deletion.
6. There are no common sequence or structural features at the sites of integration.
7. Integration appears to be the result of an illegitimate recombination event. There are no large regions of homology between viral and cellular DNA at the integration sites, although very small regions of homology (2–5 bp) may be found at virus-host joins.
8. Integration causes a rearrangement (probably a deletion) of host DNA at the integration site.
9. Initial transformants usually contain tandem arrays of viral genomes.
10. In some cases transcription initiated in viral sequences extends into host sequences, resulting in hybrid mRNAs and proteins. Processing of these mRNAs is under host control.

	VIRUS	CELL
TSA1	TGTTGGATTTCACCT	GGCTTTCCTTCCCAG
	2643	
TSA2	CTCAAGCTGAGGCAT	TTAATCCCTCTGCTC
	3833	
TSA3	CCAGCACCTCCATAC	TCTGGCTCATGCCTT
	3574	
53-RAT	ACAGGCCTAGCCGCG	GGGACAGCAGTGCCA
	2312	
82-RAT	GTACCCCAGCTCATC	AATAACTCATAACTT
	1239	
7B-L	ATGGAATGATTTCTT	GCTGTTAATTGCTAA
	916	
7B-R	TACCTACACAACGAA	TGAGTGCTATGAAAT
	4682	
7T-L	ATATCTGTATTTCCT	TTAAAAAATAATAAT
	2932	
7T-R	GAAAATGTGCCAAAA	GAACTTAAAGGGTTG
	2252	

Figure 3 The nucleotide sequence at the virus-cell joins in polyoma-transformed cell lines. The nucleotide number beneath each sequence indicates the viral nucleotide at the virus-cell join (see Fig. 2) (Hayday et al. 1982; Ruley et al. 1982, 1986).

No sequence specificity or common structural features have been detected at the virus-host joins in polyoma-transformed cells (Hayday et al. 1982; Ruley et al. 1982, 1986; Ruley and Fried 1983) (Fig. 3). The integration of polyoma sequences appears to involve nonspecific, illegitimate recombination events. No large regions of homology between viral and cellular sequences are found at the integration sites, although between 2 and 5 bp of homology may exist at the recombinant join (Stringer 1982; Ruley and Fried 1983; Williams and Fried 1986). The cellular DNA sequences on either side of the viral insert in the transformed cells are not found adjacent to each other in untransformed cells (Hayday et al. 1982; Stringer 1982), demonstrating that integration has caused a gross rearrangement (deletion or chromosomal translocation) of cellular DNA sequences at the integration site.

An inverted duplication-transposition of cellular DNA at a virus-host join

In addition to the gross rearrangement of cellular sequences at the polyoma integration site, a perturbation has been found at the left virus-host join in 7axB cells. Thirty-seven extraneous nt (filler DNA), derived from neither the viral nor directly adjacent host sequences, were found at the join when the viral sequences were compared with the cellular DNA of the unoccupied site (cellular sequences prior to the polyoma integration) (Hayday et al. 1982). By using a synthetic oligomer of these 37 nt as a probe, it was demonstrated that the filler DNA is an exact inverted duplication of a single-copy cellular sequence found 650 bp upstream from the virus-host join (Williams and Fried 1986) (Fig. 4). Thus, in addition to the large deletion or chromosomal translocation of cellular DNA at the integration site in 7axB cells, a more local inverted duplication-transposition of upstream cellular DNA has occurred, generating 37 bp

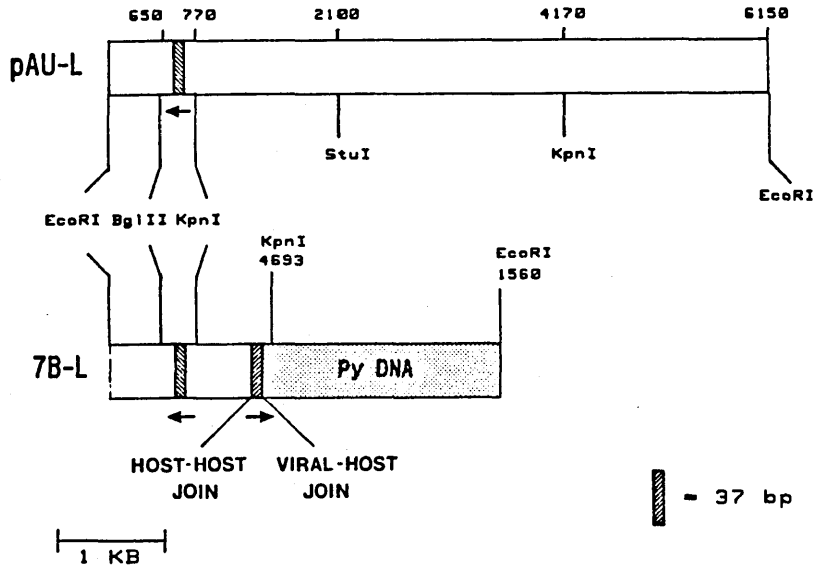


Figure 4 An inverted duplication-transposition of 37 bp at the left virus-host join in 7axB cells. Shown is a diagram of the normal upstream location of the 37 bp (hatched region) in both the unoccupied-site fragment (pAU-L) from untransformed RAT-1 cells and in the left virus-host junction fragment (7B-L) from 7axB cells. A transposed copy of the 37 bp is also found in an inverted position at the virus-host join in 7B-L. (Reprinted, with permission, from Williams and Fried 1986.)

of filler DNA at the left virus-host junction. In contrast, the right 7axB join showed an abrupt transition from viral to host sequences with 2 bp of homology at the join (Williams and Fried 1986).

Cellular sequences govern the 3′ processing of early region mRNAs initiated in the integrated polyoma DNA

Both a strong (major) and a weak (minor) polyadenylation signal are found in the polyoma early region (Fig. 1) (Kamen et al. 1980). In many polyoma-transformed cell lines, the viral sequences specifying the major polyadenylation site are lost (Fig. 2), and the vast majority of early region transcripts initiated in the viral sequences do not utilize the polyoma minor polyadenylation signal but continue past the virus-host join and are polyadenylated in the cellular sequences. In TSA-3T3 cells such viral-cellular hybrid transcripts were found to utilize a cellular polyadenylation signal (Ruley et al. 1982). The polyoma major polyadenylation signal is also absent from the transcribed early region in 53-Rat cells (Fig. 2) (Kamen et al. 1980; Ruley et al. 1986). The processed polyoma transcripts initiated in the polyoma early region of 53-Rat differ from most transcripts in other polyoma-transformed cells in that they utilize the inefficient minor polyoma polyadenylation signal instead of a cellular polyadenylation signal. Therefore the cellular DNA downstream from the polyoma insert in 53-Rat contains sequences that govern the use of the polyoma minor polyadenylation signal. These cellular sequences could contain either a transcriptional termination or processing site so that a cellular polyadenylation signal is never reached. Alternatively, transcription may proceed very slowly through these cellular sequences, allowing adequate time for the use of the inefficient minor polyoma polyadenylation signal.

Definition of the domains of the polyoma TAg's by a frameshift mutation in a polyoma-transformed cell line

Three independent flat-cell revertants isolated from the polyoma-transformed cell line 7axT (Fig. 2) have been analyzed (Wilson et al. 1986). All three revertants were observed to spontaneously retransform at a high frequency. All the revertants produced two novel TAg species of 43K and 37K instead of the 56K polyoma middle TAg and the 75K truncated large TAg synthesized by the parental 7axT cells and the spontaneous retransformants (Fig. 5). All the revertant and transformed cells also produced the polyoma 22K small TAg. A number of properties of the 43K and 37K TAg's are compared with those of polyoma large and middle TAg's in Table 2. The 43K protein shares a number of properties with the polyoma oncogene middle TAg, including its associated protein kinase activity and its ability to form a complex with c-*src*. The 43K protein differs from middle TAg in being unreactive with an anti-peptide serum, termed MTC, directed against the carboxyl ter-

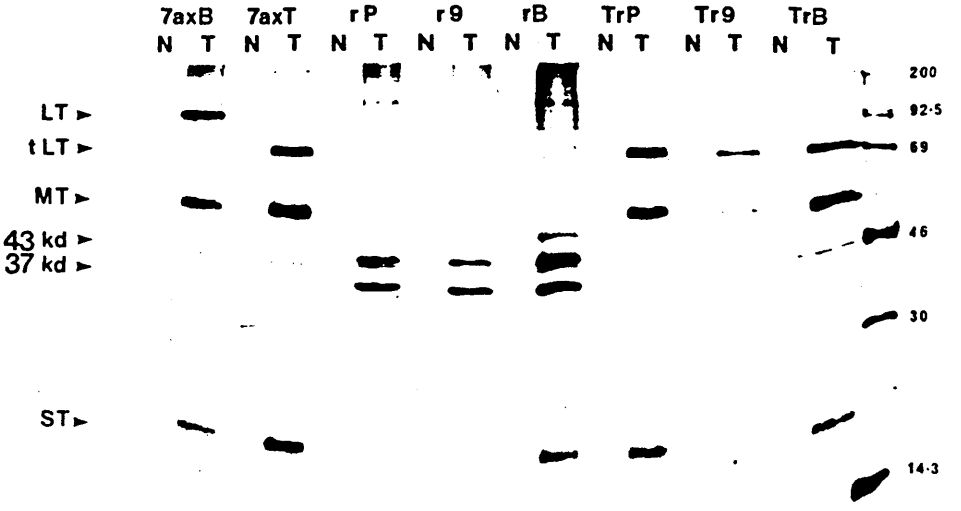


Figure 5 TAg's in transformed and revertant cell lines. Immunoprecipitation of extracts of the control polyoma-transformed cell line 7axB, the polyoma-transformed parental cell line 7axT, the three flat-cell revertants (rP, r9, and rB), and their three spontaneous retransformants (TrP, Tr9, TrB) with polyoma anti-T (T) or normal (N) serum. The positions of the following TAg's are indicated: large TAg (LT), middle TAg (MT), small TAg (ST), the 75K truncated large TAg of 7axT and the retransformants (tLT), and the novel 37K and 43K TAg species of the flat-cell revertants. (Reprinted, with permission, from Wilson et al. 1986.)

Table 2 Comparison of Properties of the Normal TAg's and the Two Truncated Proteins

Property	LT	37K	MT	43K
In vitro phosphate acceptor	–	–	+ + +	+
Association with c-*src*	–	–	+	+
In vivo phosphorylation	+ + +	+ +	+	–
Antibody reactivity:				
monoclonal LT1	+	+	–	–
monoclonal MT16	–	–	+	+
anti-peptide Glu4	–	–	+	+
monoclonal LT4	+	–	–	–
anti-peptide MTC	–	+	+	–
Location*	N	M	M	C

Data taken from Wilson et al. (1986).
*(N) Nucleus; (M) membrane; (C) cytoplasm.

minus of middle TAg, in being predominantly located in the cytoplasm instead of being membrane-associated, and in being insufficient to induce a transformed phenotype. The 37K protein resembles the polyoma large TAg in a number of respects but differs from large TAg in being unreactive with the anti-large TAg monoclonal LT-4, reacting with the middle TAg anti-peptide serum MTC and in being predominantly membrane-associated instead of being localized in the nucleus. The generation of the two novel proteins is the result of a frameshift mutation created by a deletion of one of nine consecutive cytidine residues between polyoma nt 1239 and 1247 (Fig. 6). In this region there are two open reading frames; one coding for the carboxyl terminus of middle TAg and the other for the central region of large TAg. The frameshift mutation results in the shuffling of different domains of the polyoma TAg's, generating the

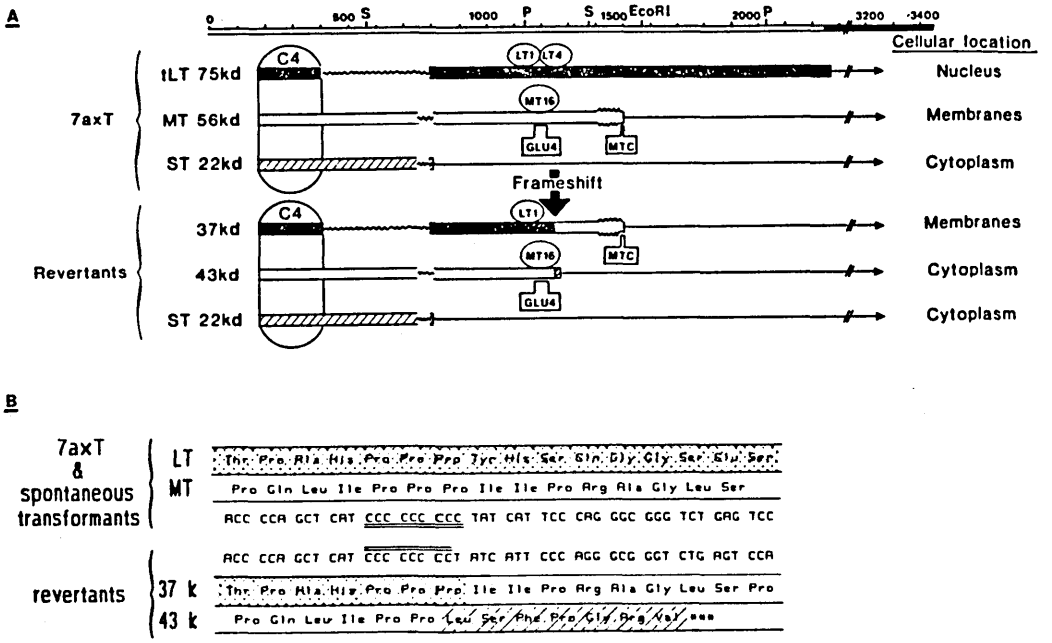


Figure 6 *Generation of the 37K and 43K novel TAg's by a frameshift at a mutational hot spot in the polyoma early region sequences encoding the carboxyl terminus of middle TAg and the central region of large TAg. (A) At top is shown the integrated polyoma early region (open band) and adjacent cellular DNA sequences (filled band) in 7axT. Below this are shown the proteins produced by 7axT, indicating the epitopes recognized by the monoclonal antibodies and the antipeptide sera specific for the polyoma TAg's. Beneath this is shown the generation of the revertant cell 37K and 43K novel TAg's as a result of frameshift mutation caused by the deletion of any one of nine cytidines between nt 1239 and 1247 (see B). This results in the large TAg reading frame (filled bands in A, cross-hatched bands in B) switching into the middle TAg reading frame (open bands) to generate the 37K protein and the middle TAg reading frame switching into the small TAg reading frame (hatched band) to generate the 43K protein (see B). (B) The deletion of a cytidine residue at a mutational hot spot in the polyoma early region causes a frameshift, generating the flat-cell revertant phenotype. The top half of the figure shows the large TAg and middle TAg reading frames above the DNA sequence of 7axT (which is the same as the sequence of the spontaneous retransformants). The bottom half of the figure shows the DNA sequence of the flat-cell revertants, which is generated by the deletion of any one of the nine consecutive cytidines between polyoma nt 1239 and 1247 (underlined). This results in the frameshift that generates the revertant 37K and 43K novel TAg's shown beneath the sequence (see A). The 37K TAg contains the amino terminus of large TAg and the carboxyl terminus of middle TAg, and the 43K TAg contains the amino terminus of middle TAg and seven amino acids at its carboxyl terminus (diagonal hatching) from the small TAg reading frame (after the termination of the small TAg protein). (Reprinted, with permission, from Wilson et al. 1986.)*

37K hybrid protein with the amino terminus of large TAg and the carboxyl terminus of middle TAg and the hybrid 43K protein with the amino terminus of middle TAg and a carboxyl terminus encoded by a normally untranslated region of the small TAg reading frame (Fig. 6). These results indicate that the hydrophobic carboxyl terminus of middle TAg is responsible for membrane association even when attached to the amino terminus of large TAg, which is thought to contain a nuclear localization signal (Richardson et al. 1986). In addition, these results show that the aminoterminal 75% of middle TAg is sufficient for its associated protein kinase activity and the ability to form a complex with c-*src* but is not sufficient to induce a transformed phenotype. Thus, some other function of intact middle TAg is required for transformation.

The possible use of frameshifting for the generation of other polyoma early region proteins

The spontaneous retransformants all contain an identical back mutation, which is a precise correction of the revertant mutation by the addition of a cytidine residue, regenerating nine cytidine residues between polyoma nt 1239 and 1247 (Fig. 6). As well as the previously described spontaneous retransformants (Wilson et al. 1986), nine other spontaneous retransformants (three derived from each of the three revertants) show the same phenotype and polyoma proteins as 7axT and presumably also contain nine cytidines in this region. The high frequency of generation of spontaneous retransformants (3×10^{-5}) must be due, at least in part, to the ability to precisely correct the revertant mutation by the addition of a cytidine residue at any one of nine different positions in the sequence (Fig. 6). The ease of isolation of deletions (revertants) and additions (spontaneous retransformants) at the cytidine run indicates that it is a hot spot for frameshift mutations. Consecutive runs of a single nucleotide have been found to be hot spots for frameshift mutations in both prokaryotes (Streisinger et al. 1967; Okada et al. 1972; Pribnow et al. 1981) and recently in eukaryotes (Baumann et al. 1985). Additions or deletions presumably arise by slippage or stuttering of the DNA polymerase at the homonucleotide run or by misalignment of the replicating and parental DNA strands, as proposed by Streisinger and Owen (1985).

In a number of organisms, a low level of normal expression from frameshift mutants has been observed, resulting in "leaky" phenotypes (Atkins et al. 1972, 1983; Fox and Weiss-Brummer 1980). There is also accumulating evidence that certain organisms, particularly viruses, utilize translational frameshifting during normal expression of their intact genomes and that this may even have a regulatory function (Kastelein et al. 1982; Dunn and Studier 1983; Craigen et al. 1985; Mellor et al. 1985; Varmus 1985). The frameshifting in viruses may be achieved with the normal translational machinery, unlike the mechanism that exists in yeast and bacteria, which relies on suppressor tRNAs containing mutant anticodons (Roth 1981). Another possible mechanism to produce transient frameshifting would be at the transcriptional level. Certain DNA sequences may cause the RNA polymerase to generate addition or deletion errors in the mRNAs, in the same manner as frameshift mutations arise during DNA replication.

It is interesting to speculate that the region of polyoma containing nine consecutive cytidines, which is a frameshift mutational hot spot for DNA replication, might also cause similar errors during transcription and/or translation of the mRNA. The ability of portions of the polyoma early region to encode hybrid proteins may endow the virus with greater genetic flexibility than would be achieved by encoding just two proteins (large and middle TAg's) in the same region of the DNA sequence. The exact polyoma gene products involved in the induction of the polyoma tumor-specific transplantation antigen (Tooze 1981) or the immortalization of primary cells (Rassoulzadegan et al. 1983) remain unclear and could be mediated by proteins made in small quantities as a result of frameshifting at either the translational or transcriptional level. Minor TAg species (37–43K) have been noted by others (Hutchinson et al. 1978; Schaffhausen et al. 1978; Silver et al. 1978; Simmons et al. 1979; Ito and Spurr 1980) during the lytic cycle and in transformed cells. In one case a 39K TAg species appeared to contain the aminoterminal amino acid sequence from large TAg and the carboxyterminal amino acid sequence of middle TAg (Ito and Spurr 1980), which is similar to the 37K TAg species produced by the revertants (Wilson et al. 1986). These minor TAg species could be specific gene products, breakdown products, or hybrid molecules of the polyoma middle and large TAg's caused by frameshifting.

Acknowledgments

The authors wish to thank Mrs. G. Briody for her help in the typing and preparation of parts of this manuscript.

References

Atkins, J.F., D. Elseviers, and L. Gorini. 1972. Low activity of β-galactosidase in frameshift mutants of *Escherichia coli*. *Proc. Natl. Acad. Sci.* **69**: 1192.

Atkins, J.F., B.P. Nichols, and S. Thompson. 1983. The nucleotide sequence of the first externally suppressible-1 frameshift mutant, and of some nearby leaky frameshift mutants. *EMBO J.* **2**: 1345.

Basilico, C., S. Gattoni, D. Zouzias, and G. Della Valle. 1979. Loss of integrated viral DNA sequences in polyoma transformed cells is associated with an active viral A function. *Cell* **17**: 645.

Basilico, C., D. Zouzias, G. Della Valle, S. Gattoni, V. Colantuoni, R. Fenton, and L. Dailey. 1980. Integration and excision of polyomavirus genomes. *Cold Spring Harbor Symp. Quant. Biol.* **44**: 611.

Baumann, B., M.J. Potash, and G. Kohler. 1985. Consequences of frameshift mutations at the immunoglobulin heavy chain locus of the mouse. *Embo J.* **4**: 351.

Cogen, B. 1978. Virus-specific early RNA in 3T6 cells infected by a *ts* mutant of polyomavirus. *Virology* **85**: 222.

Courtneidge, S.A. and A.E. Smith. 1983. Polyomavirus transforming protein associated with the product of the c-*src* cellular gene. *Nature* **303**: 435.

———. 1984. The complex of polyomavirus middle-T antigen and $pp60^{c\text{-}src}$. *EMBO J.* **3**: 585.

Craigen, W.J., R.G. Cook, W.P. Tate, and C.T. Caskey. 1985. Bacterial peptide chain release factors: Conserved primary structure and possible frameshift regulation of release factor 2. *Proc. Natl. Acad. Sci.* **82:** 3616.

Dunn, J.J. and F.W. Studier. 1983. Complete nucleotide sequences of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166:** 477.

Eckhart, W., M.A. Hutchinson, and T. Hunter. 1979. An activity phosphorylating tyrosine in polyoma T-antigen immunoprecipitates. *Cell* **18:** 925.

Fox, T.D. and B. Weiss-Brummer. 1980. Leaky +1 and −1 frameshift mutations at the same site in yeast mitochondrial gene. *Nature* **288:** 60.

Francke, B. and W. Eckhart. 1973. Polyoma gene function required for viral DNA synthesis. *Virology* **55:** 127.

Fried, M. 1965. Cell-transforming ability of a temperature-sensitive mutant of polyomavirus. *Proc. Natl. Acad. Sci.* **53:** 486.

———. 1970. Characterization of a temperature-sensitive mutant of polyomavirus. *Virology* **40:** 605.

Hayday, A.C., F. Chaudry, and M. Fried. 1983. Loss of polyomavirus infectivity as a result of a single amino acid change in a region of polyomavirus large T-antigen which has extensive amino acid homology with simian virus 40 large T-antigen. *J. Virol.* **45:** 693.

Hayday, A., H.E. Ruley, and M. Fried. 1982. Structural and biological analysis of integrated polyomavirus DNA and its adjacent host sequences cloned from transformed rat cells. *J. Virol.* **44:** 67.

Hutchinson, M.A., T. Hunter, and W. Eckhart. 1978. Characterization of T antigens in polyoma-infected and transformed cells. *Cell* **15:** 65.

Ito, Y. and N. Spurr. 1980. Polyomavirus T antigens expressed in transformed cells: Significance of middle T antigen in transformation. *Cold Spring Harbor Symp. Quant. Biol.* **44:** 149.

Ito, Y., J. Brocklehurst, and R. Dulbecco. 1977. Virus-specific proteins in the plasma membrane of cells lytically infected or transformed by polyomavirus. *Proc. Natl. Acad. Sci.* **74:** 4666.

Kamen, R., J. Favaloro, J. Parker, R. Treisman, L. Lania, M. Fried, and A. Mellor. 1980. Comparison of polyomavirus transcription in productively infected mouse cells and transformed rodent cells. *Cold Spring Harbor Symp. Quant. Biol.* **44:** 63.

Kastelein, R.A., E. Remaut, W. Fiers, and J. Van Duin. 1982. Lysis gene expression of RNA phase MS2 depends on a frameshift during translation of the overlapping coat protein gene. *Nature* **295:** 35.

Lania, L., A. Hayday, and M. Fried. 1981. Loss of functional large T-antigen and free viral genomes from cells transformed *in vitro* by polyomavirus after passage *in vivo* as tumor cells. *J. Virol.* **39:** 422.

Lania, L., M. Griffiths, B. Cooke, Y. Ito, and M. Fried. 1979. Untransformed rat cells containing free and integrated DNA of polyoma non-transforming (hr-t) mutant. *Cell* **18:** 793.

Lania, L., D. Gandini-Attardi, M. Griffiths, B. Cooke, D. De Cicco, and M. Fried. 1980. The polyomavirus 100K large T-antigen is not required for the maintenance of transformation. *Virology* **101:** 217.

Mellor, J., S.M. Fulton, M.J. Dobson, W. Wilson, S.M. Kingsman, and A.J. Kingsman. 1985. A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty1. *Nature* **313:** 243.

Okada, Y., G. Streisinger, J.E. Owen, and J. Newton. 1972. Molecular basis of a mutational hot spot in the lysozyme gene of bacteriophage T4. *Nature* **236:** 236.

Pribnow, D., D.C. Sigurdson, G. Gold, B.S. Singer, and C. Naroli. 1981. rII cistrons of bacteriophage T4 DNA sequence around the intercistronic divide and positions of genetic landmarks. *J. Mol. Biol.* **149:** 337.

Rassoulzadegan, M., Z. Naghashfar, A. Cowie, A. Carr, M. Grisoni, R. Kamen, and F. Cuzin. 1983. Expression of the large T protein of polyomavirus promotes the establishment in culture of "normal" rodent fibroblast cell lines. *Proc. Natl. Acad. Sci.* **80:** 4354.

Richardson, W.D., B.L. Roberts, and A.E. Smith. 1986. Nuclear location signals in polyomavirus large-T. *Cell* **44:** 77.

Roth, J. 1981. Frameshift suppression. *Cell* **24:** 601.

Ruley, H.E. and M. Fried. 1983. Clustered illegitimate events in mammalian cells which involve very short sequence homologies. *Nature* **304:** 181.

Ruley, H.E., L. Lania, F. Chaudry, and M. Fried. 1982. Use of a cellular polyadenylation signal by viral transcripts in polyomavirus transformed cells. *Nucleic Acids Res.* **10:** 4515.

Ruley, H.E., F. Chaudry, L. Lania, M. Griffiths, and M. Fried. 1986. Correlation of the organization and expression of the integrated polyomavirus DNA in transformed 82-Rat and 53-Rat cells. *J. Virol.* (in press).

Schaffhausen, B.S. and T.L. Benjamin. 1979. Phosphorylation of polyoma T-antigens. *Cell* **18:** 935.

Schaffhausen, B.S., J.E. Silver and T.L. Benjamin. 1978. Tumor antigen(s) in cells productively infected by wild-type polyomavirus and mutant NG-18. *Proc. Natl. Acad. Sci.* **75:** 79.

Silver, J., B. Schaffhausen, and T. Benjamin. 1978. Tumor antigens induced by nontransforming mutants of polyomavirus. *Cell* **15:** 485.

Simmons, D.T., C. Chang, and M.A. Martin. 1979. Multiple forms of polyomavirus tumor antigens from infected and transformed cells. *J. Virol.* **29:** 881.

Smith, A.E., R. Smith, B.E. Griffin, and M. Fried. 1979. Protein kinase activity associated with polyomavirus middle T-antigen *in vitro*. *Cell* **18:** 915.

Streisinger, G., Y. Okada, J. Emrich, J. Newton, T. Tsugita, E. Terzaghi, and M. Inouye. 1967. Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31:** 77.

Streisinger, G. and J.E. Owen. 1985. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. *Genetics* **109:** 633.

Stringer, J.R. 1982. DNA sequence homology and chromosomal deletion at a site of SV40 DNA integration. *Nature* **296:** 363.

Tooze, J., ed. 1981. *Molecular biology of tumor viruses*, 2nd edition, revised: *DNA tumor viruses*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Treisman, R.H., V. Novak, J. Favaloro, and R. Kamen. 1981. Transformation of rat cells by an altered polyomavirus genome expressing only the middle T protein. *Nature* **292:** 595.

Varmus, H.E. 1984. The molecular genetics of cellular oncogenes. *Annu. Rev. Genet.* **18:** 533.

———. 1985. Reverse transcriptase rides again. *Nature* **314:** 583.

Williams, T. and M. Fried. 1986. An inverted duplication-transposition event in mammalian cells at an illegitimate recombination join. *Mol. Cell. Biol.* **6:** 2179.

Wilson, J.B., A. Hayday, S. Courtneidge, and M. Fried. 1986. Frameshift in polyoma early region generates two new proteins that define T-antigen functional domains. *Cell* **44:** 477.