Volume 12 Number 11 1984 Nucleic Acids Research

Transcription from SV40-like monkey DNA sequences

Jeffrey D.Saffer and Maxine F.Singer

Laboratory of Biochemistry, National Cancer Institute, Bethesda, MD 20205, USA

Received 23 January 1984; Revised and Accepted 8 May 1984

ABSTRACT

The ability of an African green monkey genomic segment, homologous to the regulatory region of SV40, to promote transcription in monkey cells has been investigated. Genomic transcripts from CV-1 cells hybridize to both strands of this SV40-like segment. Further examination of the promoter potential of the SV40-like segment using expression vectors suggests that the SV40-like region contains at least part of the necessary information for promoting transcription in both directions. In addition, the latter experiments suggest that sequences several hundred nucleotides away in the genome modulate the transcription initiating at multiple sites in the SV40-like segment.

INTRODUCTION

Several DNA sequences that contribute to the initiation of transcription by RNA polymerase II in eukaryotes have been identified. The promoter region is usually characterized by the presence of a TATA box (consensus sequence, TATAAAA) approximately 30 nucleotides upstream of the RNA initiation site (see 1 for a review). This region is essential for efficient and precise transcription of most genes *in vitro* (2 - 5). While not essential *in vivo*, the TATA box does appear to specify a precise site of initiation (6, 7). The only examples of genes that are transcribed by RNA polymerase II but do not have a TATA box at the characteristic location are the late genes of SV40 and polyoma, the EII and IVa2 genes of adenovirus (8), and probably the U1 (9) and U2 snRNA genes (S. Van Arsdell and A. Weiner, manuscript submitted).

A second region important for RNA polymerase II transcription occurs approximately 80 nucleotides upstream of the RNA initiation site (10, 11); this region includes the sequence CAAT and has been implicated in regulation from its relatively constant location upstream of cap sites. An essential element for SV40 transcription is a G-rich octamer of the form GGGCGGPuPu which is considered a bidirectional interaction site for transcriptional factors (12- 16); in SV40 six copies of this sequence occur just upstream of the TATA box. DNA sequences even further upstream of both viral (7, 14, 17,

18) and cellular (19) genes enhance the level of transcription by RNA polymerase II without affecting the site of initiation. Such elements also enhance the transcription of genes in expression vectors (20 - 22) and have the unusual feature of functioning even if inverted (14, 18, 19). Enhancer sequences exhibit a host specificity (23 - 25) and may act as an entry site for specific factors involved in transcription (26).

We have previously isolated three different SV40-like segments from a monkey genomic library by their ability to anneal with the control region around the origin of replication of the SV40 genome (27). The homologous regions in each were subcloned and their sequences determined (27, 28). Subsequent screenings using one of these genomic segments as a probe indicated that the three SV40-like segments are members of an interspersed repeat sequence family of about 80 members (J. Saffer, unpublished results). The present experiments analyze transcription from one of these SV40-like regions called 7.02. We have detected cellular transcripts that hybridize to both strands of this region. We also demonstrate that transcription is initiated in either direction when this region is included in expression vectors and transfected into monkey cells. The potential promoter function of another SV40-like segment, 9.32 (27), was previously assessed using the same system (S. Subramani, personal communication) and with similar results. These SV40-like regions appear to be RNA polymerase II promoters although they differ from other common polymerase II promoters in structure and instead resemble, in part, the SV40 late promoter, particularly in the absence of a TATA box and the presence of the G-rich repeats.

MATERIALS AND METHODS

Constructions-All enzymes were from New England Biolabs and were used according to the purveyor's instructions. The construction of the expression vector $pA_ngpt.LR$ is shown in Fig. 1. pSV0gpt.LR was obtained from M. Fromm and P. Berg (29) and pST10 was kindly given by T. Vogel. Each of these plasmids were digested with the restriction endonucleases indicated in the Figure and the cut sites were made blunt by filling in with the Klenow fragment of *E. coli* polymerase I. Following a second digestion with Hind III, the appropriate fragments were ligated to produce $pA_ngpt.LR$. The Bam HI to Hind III fragment of pST10 contains 237 nucleotides of SV40 DNA (strain 776) in which the polyadenylylation signals for viral transcription are located. In addition, this fragment contains 18 bp of *E. coli* DNA that contribute the Hind III site. The orientation of this fragment in $pA_ngpt.LR$ is such that

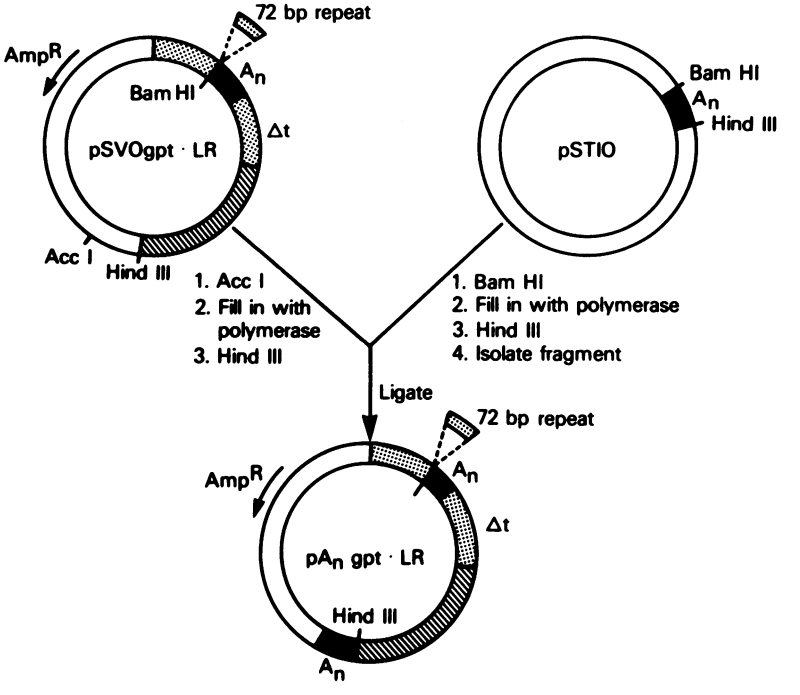


Fig. 1 Construction of $pA_ngpt.LR$. The prototype expression vector used for the experiments described was constructed as diagrammed (see Materials and Methods). The restriction sites used are noted. The unlabeled regions of pSVOgpt.LR and $pA_npt.LR$ are pBR322 sequences. The hatched area is the *E. coli* XGPRTase (gpt) gene. The stippled and black regions are segments from SV40. A_n represents the SV40 polyadenylylation signal (Bam HI to Bcl I restriction fragment). "72 bp repeat" represents nucleotides 92-272 of wild type SV40, and Δt is a fragment that contains the splice site for the small t-antigen gene.

transcripts from within the pBR322 sequences would encounter the late polyadenylylation signal before proceeding in the direction of the XGPRTase gene.

The segment 7.11 was constructed by combining the subcloned fragments 7.02 (28) and 7.06 (30) in pBR322 and recloning. Segment 7.12 is the Ava I to Acc I fragment from 7.11 and was obtained using an Acc I partial digest to generate the indicated fragment (see Fig. 6); it was purified by recloning. 7.13 is an Alu I to Sal I fragment from the previously subcloned 7.01 segment (27). Each segment - 7.02, 7.11, 7.12, and 7.13 - was inserted into the vector $pA_ngpt.LR$ using Hind III linkers. Since the vector containing the 7.12 segment showed interesting transcriptional behavior (see Results) and had been derived via

intermediate subclones, its primary nucleotide sequence was determined (31) to verify its structure. Constructions in which the SV40 72 bp repeat is deleted from the expression vector were made by digesting the appropriate vectors with Bam HI and religating.

Tissue culture and transfections- CV-1 cells were maintained in Dulbecco's Modified Eagle Medium containing 4.5 g/l glucose and no pyruvate and supplemented with 5% fetal calf serum, penicillin, and streptomycin. COS 1 cells (32) were grown in Ivcove's Modified Eagle's Medium with 5% fetal calf serum, penicillin, and streptomycin. Transfection of cells with the expression vectors was by the Ca/PO_4 method (33) with glycerol shock (34); 7.5 μg of plasmid was used without carrier for 3×10^5 cells in a 25 cm^2 dish. After 48 hr, RNA was isolated from transfected cells for S1 mapping. Estimation of the number of transformed cell colonies that express xanthine-guanine phosphoribosyl transferase was as descibed by Mulligan and Berg (35); cells were transferred to the selective medium 48 hr after transfection.

RNA mapping- Cytoplasmic RNA was isolated by the method of Favoloro et al. (36) and purified by oligo(dT) selection (37). The DNA probes were labeled at one end using [γ-^{32}P]ATP and polynucleotide kinase, and purified on a denaturing polyacrylamide gel. Hybridizations were carried out in 80% formamide at 49° according to Berk and Sharp (38); S1 nuclease digestion was at 37° for 45 min. The products of S1 digestion were analyzed on 6% acrylamide - 7 M urea gels containing Tris-borate, pH 8.3 and ethylenediaminetetraacetate.

RESULTS

Characteristics of the 7.02 SV40-like region and surrounding genomic sequences in λCaOri7-The structure of the 17.5 kb monkey segment cloned in λCaOri7 is shown schematically in Fig. 2. All of the sequences that hybridize to the SV40 probe are contained between the Sal I and Ava I endonuclease sites; this fragment was subcloned as p7.02 and its sequence has been described (28). The 17.5 kb λCaOri7 segment also contains a member of the KpnI family of long interspered repeated sequences (39) and at least three members of the Alu family (40), two of which flank the SV40-like sequence. Preliminary evidence using HeLa cell extracts in vitro (M. Lerman, unpublished results) and injection into Xenopus oocytes (S. Adeniyi-Jones, personal communication) shows that these flanking Alus are transcribed; the direction of transcription is indicated by arrows in Fig. 2. The Alu sequence to the left of the SV40-like segment represents a subclass of Alu that contains a potential Z-DNA

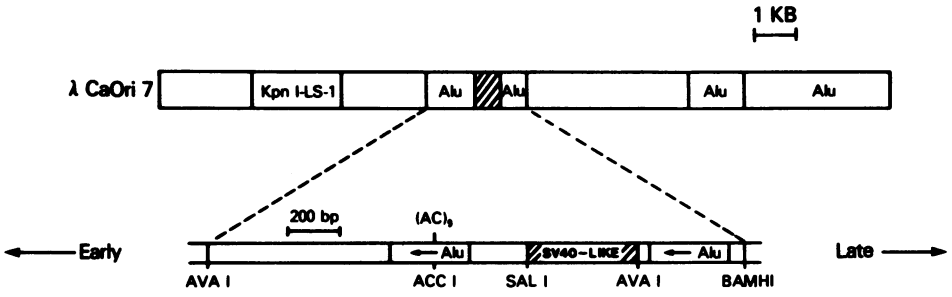


Fig. 2 Map of the monkey segment cloned in λCaOri7. The 17.5 kb insert is shown. The SV40-like segment, 7.02, is indicated by cross hatch. Alu marks restriction fragments that contain contain members of the Alu family (30, 40) and KpnI-LS1 shows the location of a member of the KpnI family of long interspersed repeated sequences (39). Unmarked regions of the map represent low copy number genomic regions. An expanded region of the map is shown below, demonstrating the relative locations of the SV40-like segment to the neighboring Alu sequences. The arrows within the Alu family members indicate the direction of transcription by RNA polymerase III. $(AC)_9$ shows the location of the potential Z-DNA forming sequence contained in the leftward Alu. The arrows labeled "early" and "late" orient the sequence in comparison to SV40. Thus, the strand going 5' to 3' from left to right contains the G-rich strand of 7.02 and is related to the late sense strand of SV40 (see Fig. 3).

forming sequence $(dA\text{-}dC)_9$ (30). The unmarked regions in Fig. 2 are of low genomic copy number, less than 1000 copies per haploid genome, as determined by hybridization of ^{32}P-labeled total genomic DNA to digests of the phage DNA. However, some regions of the 4 kb area to the right of the SV40-like segment, including 1.3 kb immediately following the Bam Hl restriction site, are known to be represented only once or a few times in the genome (J. Saffer, S. Lord, and M. Lerman, unpublished results).

The sequence (28) of the 7.02 SV40-like monkey segment studied here as well as some additional flanking sequences (30, 40) are presented together in Fig. 3. The SV40-like monkey region called 7.02 extends from nucleotides -315 to 119. Notably absent from this segment is any AT-rich stretch that is close to a canonical TATA box. The segment contains two tandem internal repeats 133 and 134 bp long, respectively, that are almost 90% homologous. The various elements of homology with SV40 in 7.02 are not colinear with the corresponding viral sequences (28). We note the presence of two pairs of direct repeats, 8 and 10 bp long that flank the SV40-like region. The two Alu sequences also form flanking, but imperfect, direct repeats. Flanking direct repeats are generally associated with the insertion of mobile elements (41, 42), but their significance here is unknown.

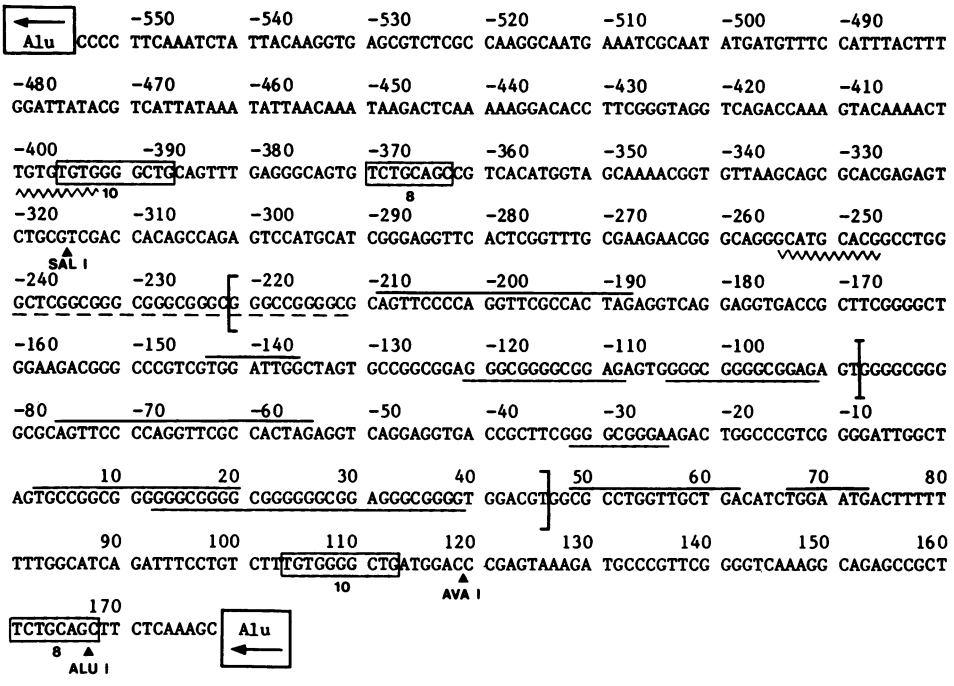


Fig. 3 The sequence of the SV40-like region, 7.02, with flanking sequences. The left to right direction is the same as that shown in Fig. 2. The sequence presented is a composite of three previously published sequences (28, 30, 40) and is read 5' to 3'. The SV40-like segment is contained between the indicated Sal I and Ava I endonuclease sites. The Alu I site shown was used for the construction of segment 7.13 (see Materials and Methods). The flanking Alu family members are indicated by labeled boxes with the arrows showing their directions of transcription by RNA polymerase III. The sequence shown extends directly up to the Alu family members and the sequence of both Alus is known (30, 40). Pairs of perfect direct repeats of 8 and 10 bp that approximately flank the region that hybridizes to SV40 are marked by boxes labeled 8 and 10, respectively. Nucleotide +1 represents the start site of the major cellular transcript in the late direction (this paper). The brackets enclose direct repeats of 133 and 134 bp. The dashed underline shows the location of the most significant homology of the 7.02 sequence to the palindrome at the SV40 origin of replication. Overlines indicate homologies to the SV40 72 bp repeats, and underlines show the location of copies of the sequence GGGCGGPuPu. Wavy underlines indicate regions of alternating purine/pyrimidine that may be capable of forming Z-DNA structures. This sequence contains four corrections from previously published data (28) at nucleotides -202, -66, 50, and 90.

We refer to the two orientations of this sequence as "early" or "late" with reference to the direction of the G-rich repeats in SV40 DNA. Thus the strand presented in Fig. 3 is the late sense strand.

Detection of cellular transcripts homologous to the SV40-like segment- Preliminary experiments showed that RNA from monkey liver and CV-1 and COS 1 cells hybridizes to the SV40-like monkey segments indicating that at least some of the 80 family members are transcribed in those cells. To investigate whether transcription of the 7.02 SV40-like sequence occurs, S1 mapping experiments were carried out. A probe that detects transcription in the late direction was prepared as indicated at the left in Fig. 4; it was labeled with ^{32}P at the Bam HI site and extends through the SV40-like sequence to the Sal I site (Fig. 2). Similarly, a probe that detects transcription in the early direction was prepared as shown on the right side of Fig. 4. It was labeled with ^{32}P at the Sal I site and extends through the SV40-like region to the Ava I site.

Hybridization of the late probe to cytoplasmic oligo(dT)-selected RNA from CV-1 cells gave rise to several specific bands after treatment with S1 nuclease (Fig. 4, left). The results imply that the 7.02 SV40-like segment is transcribed in the late direction and that at least some of the transcripts preserve the Alu after transfer to the cytoplasm. The probe used is specific for the sequence represented in λCaOri7 and should not recognize other family members since 1) the end label is in a region of low copy number, 2) it is known that not all SV40-like elements adjoin Alu sequences (27, 28), and 3) the individual members of the SV40-like family are divergent (28). The increased intensity seen on the autoradiogram at the position of the full length probe (Fig. 4, lanes 1 and 2) represents nonspecific protection of the probe by increasing amounts of RNA.

For the early direction, the probe labeled at the Sal I site at the left boundary of the 7.02 SV40-like sequence (Fig. 4, right) hybridized with the CV-1 cytoplasmic oligo(dT)-selected RNA and was digested by S1 nuclease to one size class, suggesting the SV40-like sequences are transcribed in the early direction. However, we do not know if the Sal I site is in a region of single copy sequence, and therefore it remains possible that the resulting bands in this experiment may represent hybridization to RNA from other SV40-like sequences being transcribed in the early direction.

The S1 experiments show that the 7.02 SV40-like sequence is transcribed in the late direction and suggest that 7.02 or a related segment is also transcribed in the early direction. Moreover, the size of the bands produced

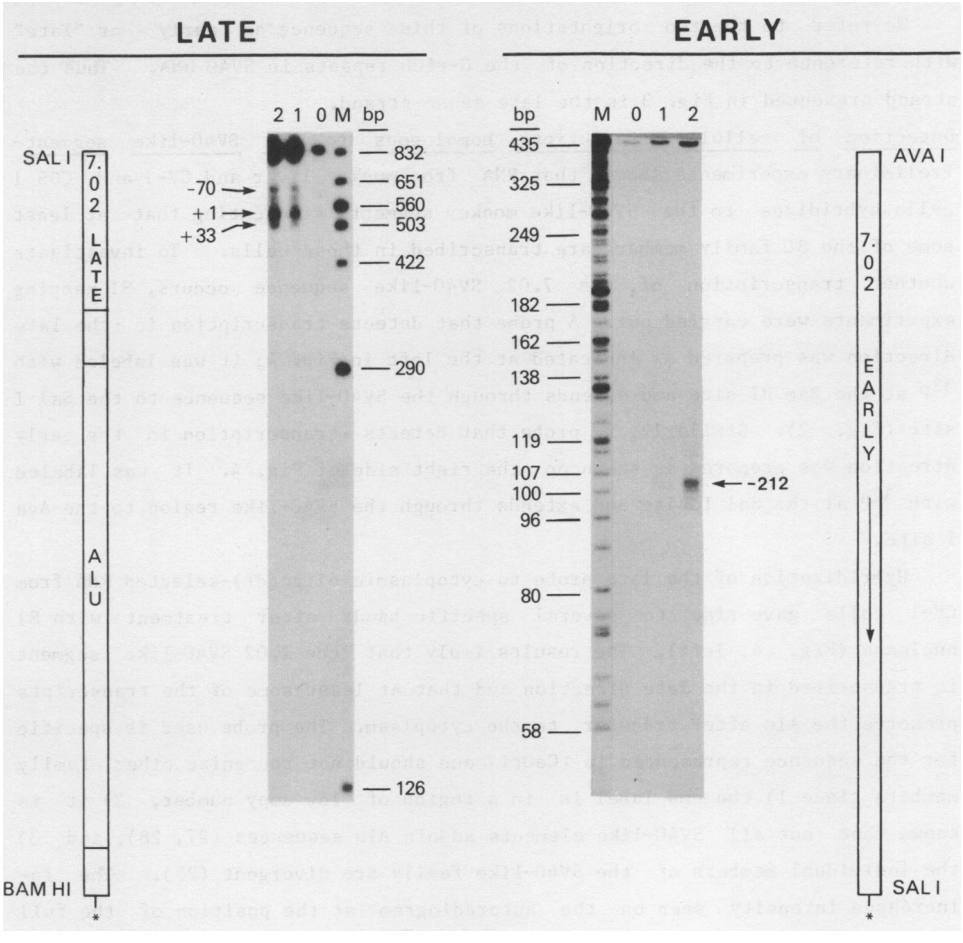


Fig. 4 S1 mapping of cellular RNA transcripts originating within the 7.02 SV40-like segment. The 5'-end labeled DNA probes shown were derived from the indicated λ CaOri7 segments and were hybridized to no RNA (lane 0), or oligo(dT)-selected cytoplasmic RNA from 6×10^6 (lane 1), or 1.4×10^7 (lane 2) CV-1 cells. The hybrids were digested with S1 nuclease (38) and the remaining DNA was analyzed on a 6% acrylamide-urea gel. The markers (M) shown were derived from the labeled probes and the sizes are shown in base pairs. The intense bands at the top of the gel represent reannealed probe. The arrows indicate the positions of the start sites and are numbered as in Fig. 3.

in the S1 experiments may correspond to the points at which transcription initiates. In the late direction, the most abundant band is at nucleotide +1 ($\pm$ 10 bp) as numbered in Fig. 3. The additional bands suggest that additional initiation sites may exist. Attempts by primer extension analysis to confirm that the bands produced with S1 are transcription initiation sites rather than

secondary processing sites were not successful. The primer was the Bam HI to Ava I fragment on the late side of 7.02 (see Fig. 2); there may be insufficient unique sequence in this segment to allow efficient competition with other Alu transcripts. In the early direction, the major S1 band corresponds to position -212 (Fig. 3).

Construction of expression vectors- The ability of the 7.02 sequence to provide for the initiation of transcription was also tested using constructed vectors that are derivatives of the pSV2gpt vectors developed by Mulligan and and Berg (29). The monkey segment containing the SV40 homology was inserted at the 5' side of the E. coli xanthine-guanine phosphoribosyl-transferase (XGPRTase) gene and the constructions were tested for XGPRTase expression in transfected monkey cells. Two methods were employed to examine transcriptional control by the inserted monkey sequence. The first is a long term assay in which transfected cells are grown under conditions that select for those cells expressing XGPRTase and the number of transformed colonies is scored. This assay measures the number of transfected cells that produce sufficient XGPRTase to survive the selective pressure. We assume that the level of enzyme in each cell depends on the strength of the promoter driving expression as well as other cellular processes that result in an active gene product. The second method is S1 mapping of transcripts formed during transient expression of XGPRTase shortly after transfection; it affords semi-quantitative data on the relative levels of the specific messenger RNAs and indicates the RNA initiation sites.

The prototype vector used for all the experiments described here is $pA_ngpt.LR$ (Fig. 4). It was constructed by inserting, 5' to the XGPRTase gene, a Bam HI/Hind III fragment of pST10 that contains the SV40 (strain 776) polyadenylylation sites for both early and late transcription inserted into the Hind III site in pSV0gpt.LR. pSV0gpt.LR itself is the same as pSV2gpt (29) except that it 1) lacks the SV40 control region to drive expression of XGPRTase and 2) includes at the Bam HI site, a segment of SV40 DNA (nucleotides 92-272) that contains the viral enhancer (designated LR for 72 bp long repeats of Fig. 1). The polyadenylylation sites were added to minimize functional transcripts that initiate at fortuitous start sites in pBR322 sequences and continue through the XGPRTase gene. Such transcripts are known to occur even in the absence of a specific promoter; they are augmented by the viral enhancer and are translated into XGPRTase (S. Subramani, personal communication). The effectiveness of the added polyadenylylation site is demonstrated in the long term selection assay; the number of colonies arising from cells transfected

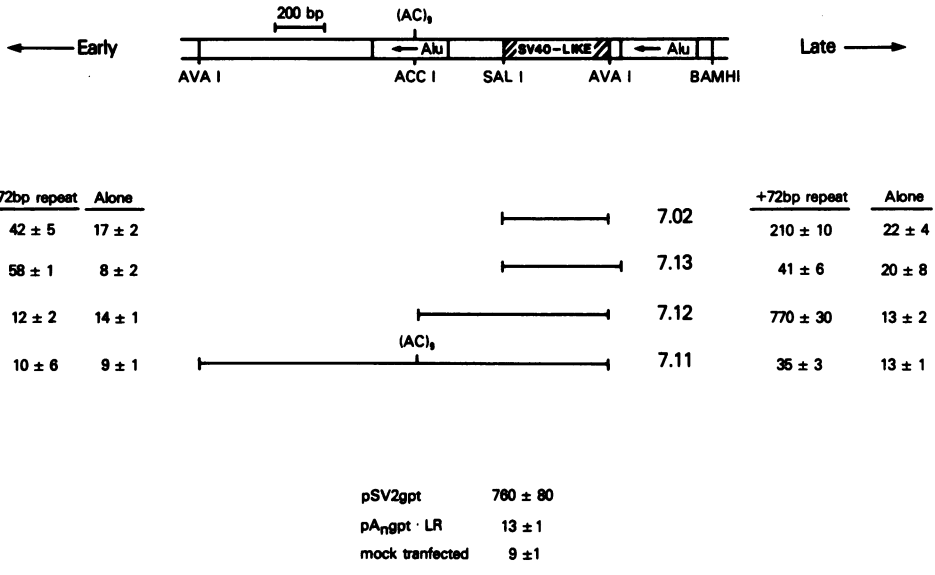


Fig. 5 Transformation of CV-1 cells after transfection with expression vectors containing the 7.02 SV40-like segment and surrounding sequences. The region of λCaOri7 from which the segments 7.02, 7.13, 7.12, and 7.11 were obtained is shown as in Fig. 2 with the location of the different subcloned segments indicated below. The numbers shown (+S.D.) represent the number of colonies visible from 10^5 transfected cells after 14 days in selective medium. The number of colonies formed after transfection with pSV2gpt and $pA_ngpt.LR$ and after mock transfection are shown below. Each subcloned segment was tested in both ("early" and "late") orientations in the expression vector, $pA_ngpt.LR$, in the presence (+72 bp repeat) and absence (alone) of the SV40 enhancer sequence.

with the modified vector $pA_ngpt.LR$ is at background levels (13 out of 10^5 cells) while pSVOgpt.LR gives rise to a 10-fold increase.

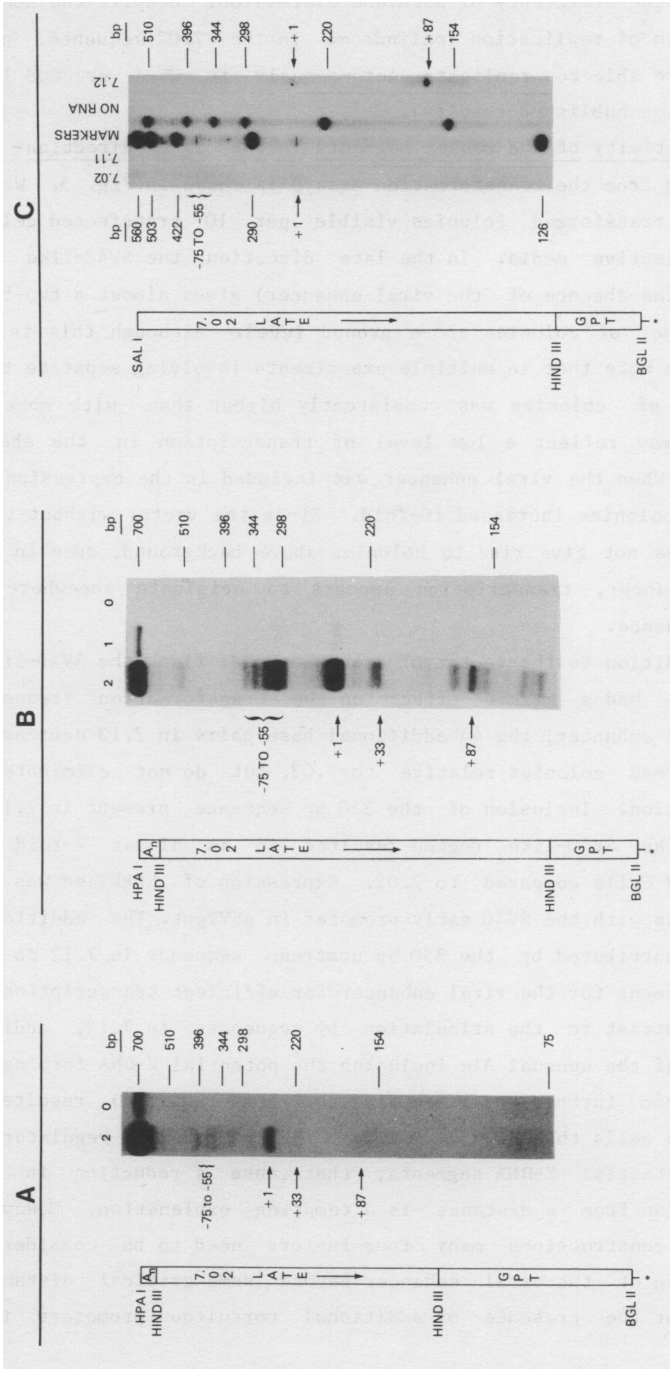
The monkey segments used in these expression vectors (Fig. 5) were inserted in both orientations into the Hind III site in $pA_ngpt.LR$ using Hind III linkers and are thus 3 kb downstream (or 2 kb upstream) from the enhancer. 7.02 is as described above; 7.13 has an additional 44 nucleotides to the right extending to within 11 bp of the flanking Alu sequence (-315 to 167 in Fig. 3). 7.12 contains the 7.02 segment plus 350 nucleotides to the left, up to but not including the $(dA\text{-}dC)_9$ segments at the junction of the two monomers in the unusual Alu. 7.11 contains the same sequence as 7.12 as well as the remainder of this Alu and 700 bp of low copy flanking sequence. Related constructions were made by deleting the SV40 enhancer. Thus we were able to test the ability of the monkey sequences to promote transcription in both directions, early and late, and to determine the effect of the viral enhancer

segment on the efficiency of XGPRTase expression. Despite the homology to the viral origin of replication palindrome in the 7.02 sequence, none of these plasmids are able to replicate autonomously in CV-1 or COS 1 cells cells (J. Saffer, unpublished results).

Promoter activity of the monkey segments in the late direction- A summary of the results from the transformation assays is shown in Fig. 5. We report the numbers of transformed colonies visible per 10^5 transfected cells after 14 days in selective media. In the late direction, the SV40-like segment 7.02 alone (in the absence of the viral enhancer) gives almost a two-fold increase in the number of colonies above ground level. Although this is not a large increase we note that in multiple experiments involving separate transfections the number of colonies was consistently higher than with mock transfected cells and may reflect a low level of transcription in the absence of the enhancer. When the viral enhancer was included in the expression vector the number of colonies increased 10-fold. Since the vector without the SV40-like segment does not give rise to colonies above background, even in the presence of the enhancer, transcription appears to originate somewhere within the monkey sequence.

The addition to the vector of sequences that flank the SV40-like region in the genome had a marked effect on the transformation frequency. In the presence of enhancer, the 44 additional base pairs in 7.13 decrease the number of transformed colonies relative to 7.02, but do not eliminate significant transformation. Inclusion of the 350 bp sequence present in 7.12 at the 5' side of the SV40-like region resulted in an almost 4-fold increase in transformed cells compared to 7.02. Expression of XGPRTase was at least as efficient as with the SV40 early promoter in pSV2gpt. The additional sequence elements contributed by the 350 bp upstream sequence in 7.12 do not replace the requirement for the viral enhancer for efficient transcription.

In contrast to the stimulation by sequences in 7.12, addition of the remainder of the unusual Alu including the potential Z-DNA forming segment and 700 bp from further upstream in the genome (7.11) resulted in fewer transformed cells than did 7.02 alone. The presence of regulatory sequences, such as potential Z-DNA segments, that cause a reduction in the level of transcription from a distance is a tempting explanation. However, in such artificial constructions many other factors need to be considered including the spacing of the viral enhancer beyond some critical distance from the promoter or the presence of additional fortuitous promoters in the extra sequence.


A
B
C
bp
700
510
396
344
298
220
154
75
560
503
422
290
126
-75 to -55
+1
+33
+87
HPA I
HIND III
SAL I
BGL II
7.02
7.11
MARKERS
NO RNA
7.12

Overall, these transformation assays demonstrate that the SV40-like sequence in the late orientation contains information that allows transcription of the XGPRTase gene. In addition, the nature of the vector suggests this transcription must initiate within the monkey segment. These conclusions are supported by mapping the RNA initiation sites on the vectors upon transient expression. The probe used for S1 mapping was end labeled at a Bgl II site 120 bp into the E. coli XGPRTase sequence and extended through the inserted monkey sequence (Fig. 6) and was thus specific for transcripts arising from the vector. In the late direction, the SV40-like segment 7.02 alone gave rise to one major band and multiple minor bands (Fig. 6A). The major band at +1 ($\pm$5 bp) corresponds to the major S1 band produced by cellular transcripts in the late direction. The secondary sites between nucleotides -75 and -55 correspond to the secondary cellular sites. The most intense band in the area of 160 nucleotides is at position +87 in the sequence (Fig. 3) and does not correspond to an observed cellular initiation site. Additional gels using Maxam and Gilbert sequencing reactions of the probe as markers were used to determine the start sites 2 bp in this and

Fig. 6 S1 mapping of transcripts produced from expression vectors containing the SV40-like sequence in the late orientation after transfection into monkey cells. A. The probe used is shown schematically at the left; it was 5'-end labeled at a Bgl II site 120 bp within the XGPRTase (gpt) gene as indicated by the asterisk and extended through the 7.02 segment to a Hpa I site within the polyadenylylation site (A_n). Hybridization with no RNA (lane 0) or oligo(dT)-selected cytoplasmic RNA from 1.4 x 10^7 CV-1 cells after transfection with $pA_ngpt.LR$ containing 7.02 in the late orientation (lane 2) was as described in Materials and Methods. The markers were ^{32}P-labeled Hinf I restriction fragments of pBR322 and the sizes are indicated in base pairs. The start sites marked by arrows are numbered according to Fig. 3. B. The probe shown at the left is the same as in part A. The RNA used for these S1 mapping experiments was oligo(dT)-selected cytoplasmic RNA from COS 1 cells transfected with $pA_ngpt.LR$ containing the 7.02 segment in the late orientation and with an entire SV40 regulatory region (Hind III to Pvu II fragment) replacing the 72 bp repeat. Hybridizations were with no RNA (lane 0) or RNA from 6 x 10^6 (lane 1) and 1.4 x 10^7 (lane 2) transfected cells. Markers are as in part A. Significant start sites are labeled with arrows or a bracket and are numbered according to Fig. 3. C. S1 mapping of RNA from CV-1 cells transfected with the expression vector containing either 7.02, 7.11, or 7.12 in the late orientation. RNA from 5 x 10^6 cells was mapped as described above using a probe (left) which was 5'-end labeled at the Bgl II site within the XGPRTase (gpt) gene and extended through the 7.02 segment to the Sal I restriction site. A control hybridization without RNA is shown in the lane labeled "no RNA". Two sets of markers, one derived from the probe and the other from pBR322 digested with Hinf I, are shown with the sizes stated in base pairs. The start sites are marked by arrows or a bracket and are numbered according to Fig. 3. The band at nucleotide 1 in lane 7.12 is partially obscured by an artifact in this exposure but is clearly visible in other experiments.

later experiments. (Confirmation of the S1 mapping bands as initiation sites rather than splicing acceptor sites was approached using primer extension. Bands were seen corresponding to each observed S1 band but in addition several shorter bands were found in the area of nucleotides 5 to 40, presumably due to termination of reverse transcription in this G-rich region.)

An additional construction was made containing the 7.02 SV40-like segment in the promoter site and the entire SV40 regulatory region instead of just the 72 bp repeat at the Bam HI site. This plasmid was transfected into COS 1 cells in which it can replicate. S1 mapping of the initiation sites on this plasmid revealed no detectable variations from the results with the nonreplicating plasmids in CV-1 cells except for the expected increase in RNA resulting from the increased copy number of the template (Fig. 6B). In principle, part of the stimulation might also come from direct interaction of T-antigen with the SV40-like region. Thus, unlike the early start sites in SV40 itself, which are shifted upon replication (14), late direction transcription from the monkey SV40-like segment was unaffected by replication.

The construction containing segment 7.12 in the late orientation that gave rise to an enhanced number of selectable colonies (Fig. 5) was also found to produce more RNA than 7.02. As seen in Fig. 6C, S1 experiments using comparable amounts of RNA show detectable transcripts for the 7.12 plasmid, even at an exposure when very little RNA is evident with 7.02 in concurrence with the difference in the number of transformed colonies in the long term assay. The major band from the 7.12 plasmid occurs at nucleotide +87 which was a minor band with 7.02. The secondary site at +1 does correspond to the major cellular and 7.02 sites. Since the hybridizations and S1 digestions were done identically to those for the other constructions, this apparent start site is not an artifact from the neighboring T-rich region. Thus the addition of upstream sequences enhanced total RNA production and shifted the distribution of sizes of the RNAs. The late construction containing 7.11 exhibits the same start as the major site in 7.12 (Fig. 6C) but at a reduced level as predicted by the long term assay (Fig. 5). In the exposure shown, the band at nucleotide +87 in the lane for 7.11 is not readily visible, but is obvious on longer exposure.

The occurrence of the major 7.12 band at a location distinct from the observed cellular transcription initiation sites raises an interesting question. Although the 7.12 segment contains more upstream sequence than does 7.02, transcription from 7.12 is less like the <u>in vivo</u> pattern than is

transcription from the shorter segment. One explanation for this observation is that because of the SV40 enhancer, which is not available in the cellular genome, normally quiescent positions serve as transcription inititation sites. Promoter activity of the monkey segments in the early direction- The results from the long term transformation assays using vectors containing the SV40-like segment in the early orientation are included in Fig. 5. Again colonies are found only when the monkey sequences are included upstream of the XGPRTase gene, implying that transcription initiates within the SV40-like segment. As seen for the late construction, the activity of 7.02 in the early orientation was stimulated by the viral enhancer. The monkey segment 7.13, when used in the early orientation, also depended on the presence of the viral enhancer for efficient expression and gave approximately the same number of colonies as 7.02 alone. Expression in the early direction from the segments 7.12 and 7.11 would require progression of the transcript through the additional sequences. Thus the absence of colonies in these assays may reflect premature termination of transcription, improper splicing, transcript instability, or improper translation of the longer message, rather than the absence of initiation.

S1 analysis of transcripts from 7.02 in the early direction is also characterized by multiple bands (Fig. 7A). In this orientation, the major bands correspond to nucleotides +20, -80, and -100 (Fig. 3) with transcription proceeding toward the Sal I site. These bands were the same in a replicating plasmid transfected into COS 1 cells (Fig. 7B), analogous to results with the construction using the late promoter. (As for the late orientation constructions, the S1 mapping bands were observed in primer extension experiments. However, reverse transcription of this sequence does not proceed efficiently past nucleotide -185 and, as a result, the major bands at positions +20, -80, and -100 are very weak.) The site of initiation of cellular transcripts (-212) is represented by only a minor band from within the vector. The reason for the difference is not clear, but as with the 7.12 segment in the late orientation, possible explanations include the unnatural effect of the viral enhancer on transcription or the inability of the plasmid to duplicate specific chromatin structures that occur in the cellular sequences.

DISCUSSION

We have presented data indicating that in CV-1 cells transcription comes from within the genomic SV40-like sequence 7.02 in two directions. Moreover,

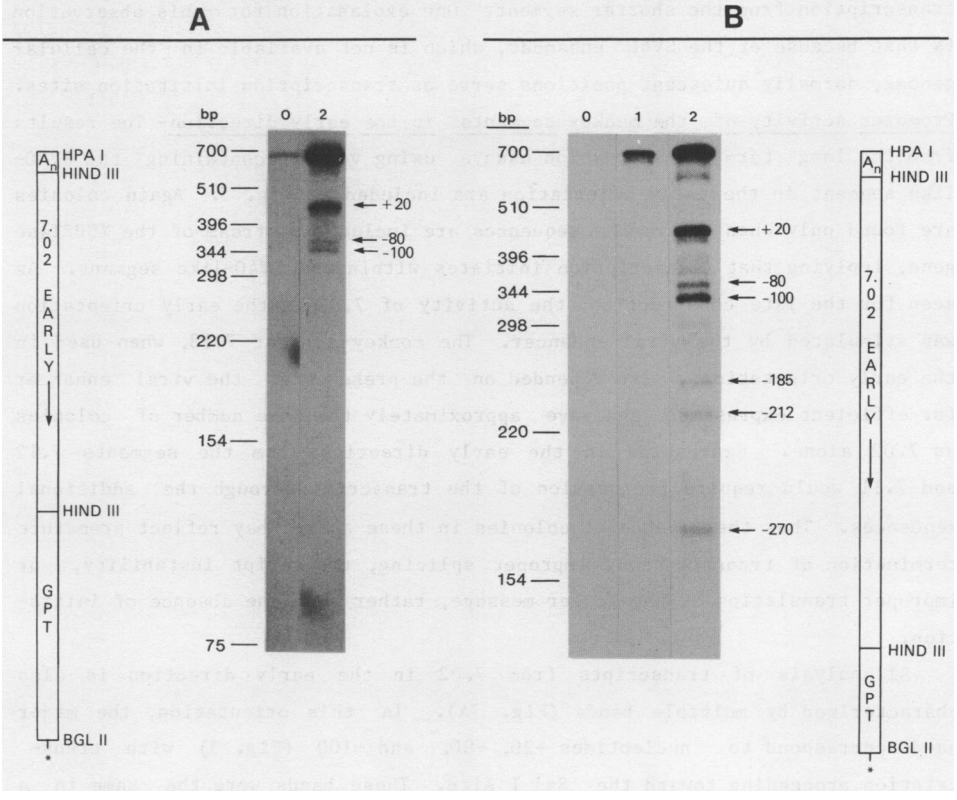


Fig. 7 S1 mapping of transcripts produced from expression vectors containing 7.02 in the early orientation after transfection into monkey cells. The probes used for hybridization, shown to the left for part A and to the right for part B, and the experimental design are analogous to those described in Fig. 6A and B. A. Mapping transcripts in CV-1 cells. The lane marked 0 contains no RNA in the hybridization and the lane marked 2 used RNA from 1.4 x 10^7 cells transfected with $pA_ngpt.LR$ containing 7.02 in the early orientation. The markers are as in Fig. 6A; the start sites are marked by arrows and are numbered according to Fig. 3. B. Mapping transcripts in COS 1 cells. Lane 0 is a control hybridization with no RNA and lanes 1 and 2 have increasing amounts of RNA as in Fig. 6B. The Hinf restriction fragments of pBR322 used as markers are marked in base pairs and the start sites are indicated as above.

regardless of its orientation, the 7.02 sequence promotes the initiation of transcription from expression vectors after transfection into homologous cells even though it lacks a TATA box. The observation of multiple apparent initiation sites may be associated with the absence of the TATA box. Similar results were obtained by S. Subramani (personal communication) using expression vectors containing another member of the family of SV40-like monkey segments. Thus the behavior of the 7.02 segment is not unique among

the family of monkey sequences that hybridize to the regulatory region of SV40. We consider the promoter activity of the SV40-like segment to be specific for RNA polymerase II; initial work in vitro using α-amanitin (M. Lerman, unpublished results) is consistent with this assumption.

The definition of 7.02 as the SV40-like region is an arbitrary designation based on fortuitous restriction endonuclease sites. The segment between the Sal I and Ava I sites contains all sequences within λCaOri7 that hybridize to the SV40 control region, but does not necessarily define an entire functional control element. In fact, the effect of upstream sequences on the position and relative efficiency of the RNA initiation sites suggests that the entire functional unit may be quite large. In addition, no expression vector used mimicked exactly initiation from the cellular sequence in vivo, emphasizing that other factors, such as chromatin structure over a large region, may be operative in the cell.

In the expression vectors, the SV40-like region depends on the presence of a viral enhancer for high levels of transcription. However, transcription from the sequence within cellular chromatin is readily detectable although no enhancer activity was evident in expression vectors in the fragments tested despite the presence of two regions within 7.02 (positions -143 to -137 and 67 to 73) homologous to the SV40 enhancer core sequence (24). This may reflect the presence of a cellular enhancer outside the segments studied. However, the role of the viral enhancer is not necessarily to replace a missing cellular enhancer; it may instead replace other distinctive control systems including specific chromatin structure. In monolayer monkey cells the SV40 enhancer replaces sequences necessary for the characteristic regulation of a Drosophila heat shock gene (43); in HeLa cells it allows correct initiation of transcription of the λI light chain gene (44), and partially compensates for the lack of the 21 bp repeats in a deletion mutant of SV40 (45). Although the expression vectors are an artificial system, the results support the in vivo findings that transcription can initiate at multiple sites in the SV40-like region in both directions and suggest the regions tested contain at least some of the promoter elements.

The SV40-like promoter region extends over several hundred nucleotides. Since transcriptional start sites are asymmetric relative to the two long tandem repeats, some of the information for start site determination is likely to be external to these regions. Alternatively, the 10% sequence divergence between the repeats may alter essential information in one. The G-rich repeats of SV40 are considered to be RNA polymerase II interaction sites (12 - 14),

and the positions of the G-rich repeats in 7.02 relative to the late initiation sites are notable. With the exception of the +87 start site, a copy of the G-rich octamer occurs approximately 30 bp upstream of each start site.

One obvious difference between 7.12 in the late direction and the cellular unit is the presence of the downstream Alu in the latter. Preliminary results using the segment 7.12 joined to the rightward Alu show a reduction in transient RNA and in colony formation in the transformation assay when compared to 7.12 alone (J. Saffer, unpublished results). As already noted, this Alu is transcribed in cell free systems; potential termination sites for this pol III transcript are between the Sal I restriction site and the leftward Alu (nucleotides -345, -403, and -438, Fig. 3). As with the monkey genomic segment, transcription of an Alu-like sequence in an intron of the mouse alpha-fetoprotein gene by polymerase III is in the opposite direction from the polymerase II transcription (46); deletion of the Alu-like sequence in a constructed minigene resulted in an approximately 5-fold increase in the level of transcripts from of the alpha-fetoprotein gene (T. Vogt and S. Tilghman, personal communication). Thus, modulation of polymerase II transcription by Alu sequences may be a general mechanism.

Classical studies on promoters have involved characterization of the sequences that promote expression of a known gene. For the SV40-like region studied here, we cannot identify a gene product nor have we yet defined the entire transcriptional unit(s). At this point we know only that cellular transcripts include both strands of the 7.02 segment. There is ample low copy number sequence in each direction to code for substantial sized proteins. But even in the absence of a known gene product, our data demonstrate that the SV40-like monkey sequences share functional as well as sequence homology to the viral control region. Most importantly, the monkey segment serves as the site of initiation for transcription in both directions and, like SV40 late transcription, it does so independent of any TATA box. The most striking structural homology between the monkey segment and the SV40 control region is the set of G-rich repeats, which are specifically implicated in viral transcription and may prove to be a critical element in transcriptional initiation in the monkey genome. These findings support the use of SV40 as a model for cellular processes.

ACKNOWLEDGEMENTS

We thank B. Howard and R. Padmanabhan for their instruction on transfections; M. Fromm, P. Berg, and T. Vogel for their generous gifts of

plasmids; the Eli Lilly Co. for the mycophenolic acid; and M. Lerman, S. Adeniyi-Jones, S. Subramani, T. Vogt and S. Tilghman, and S. van Arsdell and A. Weiner for communicating their results prior to publication.

REFERENCES

1. Shenk, T. (1981) Current Top. Micro. Immun. 93, 25-40
2. Mathis, D. J. and Chambon, P. (1981) Nature (London) 290, 310-315
3. Grosveld, G. C., Shewmaker, C. K., Jat, P., and Flavell, R. A. (1981) Cell 25, 215-226
4. Osborne, T. F., Gaynor, R. B., and Berk, A. J. (1982) Cell 29, 139-148
5. Zarucki-Schultz, T., Tsai, S. Y., Itakura, K., Soberon, X., Wallace, R. B., Tsai, M.-J., Woo, S. L. C., and O'Malley, B. W. (1982) J. Biol. Chem. 257, 11070-11077
6. Grosschedl, R. and Birnstiel, M. L. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 1432-1436
7. Benoist, C. and Chambon, P. (1981) Nature (London) 290, 304-310
8. Baker, C. C. and Ziff, E. B. (1981) J. Mol. Biol. 149, 189-221
9. Murphy, J. T., Burgess, R. R., Dahlberg, J. E., and Lund, E. (1982) Cell 29, 265-274
10. Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) Nucleic Acids Res. 8, 127-142
11. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spiritz, R. A., DeReil, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) Cell 21, 653-668
12. Myers, R. M., Rio, D. C., Robbins, A. K., and Tjian, R. (1981) Cell 25, 373-384
13. Lebowitz, P. and Ghosh, P. K. (1982) J. Virol. 41, 449-461
14. Fromm, M. and Berg, P. (1982) J. Mol. Appl. Genet. 1, 457-481
15. Everett, R. D., Baty, D., and Chambon, P. (1983) Nucleic Acids Res. 11, 2447-2464
16. Dynan, W. S. and Tjian, R. (1983) Cell 35, 79-87
17. Gruss, P., Dhar, R., and Khoury, G. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 943-947
18. Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M. P., and Chambon, P. (1981) Nucleic Acids Res. 9, 6047-6068
19. Grosschedl, R. and Birnstiel, M. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 7102-7106
20. Picard, D. and Schaffner, W. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 417-421
21. Banerji, J., Rusconi, S., and Schaffner, W. (1981) Cell 27, 299-308
22. de Villiers, J. and Schaffner, W. (1981) Nucleic Acids Res. 9, 6251-6264
23. de Villiers, J., Olson, L., Tyndall, C., and Schaffner, W. (1982) Nucleic Acids Res. 10, 7965-7976
24. Laimins, L. A., Khoury, G., Gorman, C., Howard, B., and Gruss, P. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 6453-6457
25. Byrne, B. J., Davis, M. S., Yamaguchi, J., Bergsma, D. J., and Subramanian, K. N. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 721-725
26. Wasylyk, B., Wasylyk, C., Augereau, P., and Chambon, P. (1983) Cell 32, 503-514
27. McCutchan, T. F. and Singer, M. F. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 95-99
28. Queen, C., Lord, S.T., McCutchan, T. F., and Singer, M. F. (1981) Mol. Cell. Biol. 1, 1061-1068
29. Mulligan, R. C. and Berg, P. (1980) Science 209, 1422-1427

30. Saffer, J. D. and Lerman, M. I. (1983) Mol. Cell. Biol. 3, 960-964
31. Maxam, A. M. and Gilbert, W. (1980) Meth. Enzymol. 65, 499-580
32. Gluzman, Y. (1981) Cell 23, 175-182
33. Graham, F. L. and Van der Eb, A. J. (1973) Virology 52, 456-467
34. Parker, B. and Stark, G. W. (1979) J. Virol. 31, 360-369
35. Mulligan, R. C. and Berg, P. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 2072-2076
36. Favaloro, J., Treisman, R., and Kamen, R. (1980) Meth. Enzymol. 65, 718-749
37. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, Cold Spring Harbor
38. Sharp, P. A., Berk, A. J., and Berget, S. M. (1980) Methods Enzymol. 65, 750-768
39. Lerman, M. I., Thayer, R. E., and Singer, M. F. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 3966-3970
40. Grimaldi, G., Queen, C., and Singer, M.F. (1981) Nucleic Acids Res. 9, 5553-5568
41. Mulligan, R. C. and Berg, P. (1981) Mol. Cell. Biol. 1, 449-459
42. Grimaldi, G. and Singer, M. F. (1982) Proc. Natl. Acad. Sci. 79, 1497-1599
43. Pelham, H.R.B. (1982) Cell 30, 517-528
44. Picard, D. and Schaffner, W. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 417-421
45. Hartzell, S. W., Yamaguchi, J., and Subramanian, K. N. (1983) Nucleic Acids Res. 11, 1601-1616
46. Young, P. R., Scott, R. W., Hamer, D. H., and Tilghman, S. M. (1982) Nucleic Acids Res. 10, 3099-3116