Sequences Locating the 5' Ends of the Major Simian Virus 40 Late mRNA Forms

MICHAEL PIATAK,^{1†} PRABHAT K. GHOSH,² LEONARD C. NORKIN,³ AND SHERMAN M. WEISSMAN^{1*}

Department of Human Genetics¹ and Department of Internal Medicine,² Yale University School of Medicine, New Haven, Connecticut 06510, and Department of Microbiology, University of Massachusetts, Amherst, Massachusetts 01002³

Received 21 April 1983/Accepted 4 August 1983

The 5' sequences of late mRNA specified by several constructed or naturally occurring deletion or duplication mutants of simian virus 40 were examined. The mutants included viruses with various small deletions centered about 25 nucleotides upstream from the major transcription initiation site, as well as viruses containing tandem duplications of a sequence of 50 nucleotides or less embedding the major transcription initiation site. The results show that the sequences 25 to 30 nucleotides upstream from the major initiation site in the position of the TATA box of other polymerase II promoters are not essential for the precise localization of the initiation site of late mRNA. Rather, we deduce that the major late mRNA start site is determined primarily by sequences located very close to the initiation site, and that the relative abundance of the 5' ends with this initiation site is modulated by nearby downstream sequences. Modification of six nucleotides adjacent upstream to the initiation site almost completely prevents the utilization of this site. Various deletions and substitutions of sequences 21 nucleotides or more downstream from the major initiation site causes upstream shifts in the localization of the most abundantly utilized 5' ends. The sequences immediately downstream from the major simian virus 40 initiation sites contain inverted symmetries that could give rise to secondary structures in either single-stranded DNA or RNA; the possibility that these inverted symmetries function in transcription initiation at the level of DNA structure rather than in RNA stabilization is discussed. Finally, we present additional evidence that precursor species with certain 5' termini are selectively spliced to form 19S RNA, whereas other 5' termini are preferred for forming the 16S RNA splice. We discuss the possibility that this is a consequence of the influence of leader structure on downstream splicing events.

In most eucaryotic systems, an RNA polymerase II promoter region is typified by the presence of an A+T-rich sequence of approximately eight bases located upstream from the major transcription initiation site. Commonly termed the Hogness-Goldberg box, this sequence of the form TATAAA importantly positions the initiation of major RNA transcripts +25 to +30 nucleotides downstream (2, 7, 11, 13, 35, 39), although the precise initiation site may be locally heterogeneous (13, 23). In contrast, the simian virus 40 (SV40) late transcription unit lacks any obvious Hogness-Goldberg box sequence. The locations of the 5' ends of SV40 late RNA are extremely heterogeneous, spanning some 300 nucleotides (4, 17) and are

not precisely located with respect to any particular short sequence either upstream or downstream. The locations of the 5' ends of SV40 early RNA (13) and sea urchin histone RNAs (18) show increased heterogeneity if the Hogness-Goldberg box is deleted, but not to an extent approximating that for SV40 late RNA. The analogous transcription unit in the related papovavirus, polyomavirus, also exhibits extreme heterogeneity in the location of RNA 5' ends. A Hogness-Goldberg-like box is present in this system, but only a small proportion of the abundant late RNA 5' ends is located 30 ± 5 nucleotides downstream from it (9).

Studies of the late RNA produced by SV40 mutant viruses in which DNA encoding the major late RNA 5' ends and other portions of the late leader sequence are deleted suggested that sequences downstream of the major late RNA 5' end site (14, 21, 22, 41) help specify the utiliza-

† Present address: Cetus Corp., Emeryville, CA 94608.

tion of that site. In particular, deletion mutants removing sequences 25 or more residues downstream of the major initiation site caused substantial upstream shifts in the location of the 5' ends of the majority of the late RNA (41).

In the present report we have reinvestigated certain aspects of SV40 late mRNA production *in vivo* and constructed new mutants in which segments of the putative promoter region for SV40 late RNA are either deleted or duplicated in an attempt to further define sequences involved in late RNA production. The studies suggest that the sequence adjacent to and the sequences near downstream from the initiation sites for late transcription play critical roles in locating the sites of initiation of this late transcription. A hypothetical model able to account for these influences is discussed.

MATERIALS AND METHODS

Most of the experimental procedures, including conditions for cell culture and infection, the wild-type virus used (776), nucleotide numbering system, preparation of late cytoplasmic and nuclear polyadenylated RNA, and 5'-³²P primer extension analysis of SV40-specific RNA, were as described in detail previously (41). The primer for 16S RNA is the *Hae*III-*Hind*III SV40 DNA fragment hybridizing to residues 1382 through 1415; the primer for 19S RNA is the *Sau*96I DNA fragment hybridizing to residues 479 through 509. In the present studies the CV-1 continuous line of African green monkey kidney cells was used. The virus 777N is presumably the wild-type variant 777, but is here designated by the additional letter N (Norkin) to indicate the particular isolate used. This virus and the virus SP1 have been described previously (37), and their genomes have been partially sequenced (38).

Restriction endonucleases and T4 DNA ligase were purchased from New England Biolabs, Beverly, Mass. *Escherichia coli* DNA polymerase, large fragment, and (γ -³²P)ATP and α -³²P-deoxynucleotide triphosphates were purchased from New England Nuclear Corp., Boston, Mass. S1 nuclease was purchased from Bethesda Research Laboratories, Rockville, Md. All commercial enzymes were used according to the manufacturer's conditions. T4 polynucleotide kinase was supplied by Richard Wang of this laboratory and used as described previously (41).

Construction of mutant viruses. The viruses KMTD2 and KMTD3, in which the region from the *Kpn*I site (map unit [m.u.], 0.715; residue 212) to the *Msp*I site (m.u., 0.725; residue 265) is effectively duplicated in tandem, were constructed by ligating two appropriately derived fragments of SV40 DNA (Fig. 1). Two samples of wild-type viral DNA were digested with *Kpn*I and *Msp*I, respectively. The protruding ends were then rendered blunt by a gentle digestion with 1 U of S1 nuclease per μ g of DNA for 30 min at room temperature followed by 10 min at 4°C. The reaction was stopped by adding 0.1 volume of 1 M Tris–20 mM EDTA (pH 9.4) and extracting twice with an equal volume of phenol-chloroform-isoamyl alcohol (24:24:1) containing 0.05% hydroxyquinoline and equilibrated to 50 mM Tris (pH 7.5). These linear

DNAs were then cleaved with *Eco*RI. The resulting fragments were separated in a 4% polyacrylamide gel in 40 mM Tris acetate (pH 7.8)–1 mM EDTA, eluted as described previously (15), and concentrated by ethanol precipitation. The DNA fragment containing the early portion of the genome from the blunt *Msp*I site to the *Eco*RI site was ligated to the shorter fragment containing part of the late region from the blunt *Kpn*I site to the *Eco*RI site (Fig. 1). Various amounts of the ligated material (from 0.1 to 10 ng/10⁶ cells) were transfected into CV-1 cells by the method of Chu and Sharp (5), except that the glycerol shock treatment was not done. (It is our experience that inclusion of a glycerol shock treatment results in a substantial loss in viable cells and an overall decrease in the number of transfected cells.) At 24 h after transfection, cell monolayers were overlaid with 1.5% agar containing minimal essential medium (Earle salts) supplemented with 2 mM glutamine, 250 U of penicillin G per ml, 250 μ g of streptomycin sulfate per ml, 2% agammaglobulin, and heat-inactivated calf serum, but lacking phenol red. Monolayers were fed every fourth day with the agar medium, but containing only 0.5% serum, until plaques were visible. The monolayers were then stained with neutral red (0.001% in an agar medium overlay) to facilitate visualizing the plaques. Individual plaques were picked into sterile 50 mM Tris (pH 8.0) and extracted once with chloroform; the virus was then propagated in CV-1 cells, and virus stocks were prepared as crude freeze-thawed lysates from monolayers showing about 80% cytopathic effect (CPE). Individual virus stocks were screened for genomic construction and purity by examining an appropriate restriction digest pattern of a small amount of viral DNA prepared as described below. KMTD2 and KMTD3 were isolated as having potential *Kpn*I-*Msp*I tandem duplications.

d/KMTD3/2-7 and other derivatives of the tandem sequence construction in the KMTD viruses were constructed in a fashion similar to that above (Fig. 2). The constructions differed in two important ways. First, a more extensive treatment with S1 nuclease was performed on those restriction sites around which a deletion was desired. For example, in constructing *d*/KMTD3/2-7, *Kpn*I-cleaved KMTD3 parent DNA was treated with S1 nuclease at 37°C for 30 min. The appropriate fragment of the genome was ligated to the complementary fragment of *Kpn*I-cleaved wild-type DNA made just blunt end with a brief S1 nuclease treatment to obtain an essentially unidirectional deletion of sequences downstream of the *Kpn*I site in KMTD3. Second, genome portions were not separated on polyacrylamide gels before ligation as above; ligation and selection of the appropriate genome portions were optimized by cleaving undesired regions with other restriction enzymes followed by treatment with S1 nuclease (Fig. 2). Upon ligation via the unmodified *Eco*RI ends, the desired and singularly viable combination of viral genomes was selected by transfection of CV-1 cells. This method was also applied in generating viruses (data not shown) similar to KMTD2 and KMTD3, but having tandemly duplicated segments shortened by deletions approximately midway between the remaining *Kpn*I and *Msp*I sites.

Preparation of screenable amounts of viral DNA. Small amounts of viral DNA suitable for restriction analysis were prepared as follows. One milliliter of

FIG. 1. Construction scheme for KMTD viruses.

virus stock (titer unknown) was used to infect a 10-cm culture dish of CV-1 cells. After 65 to 72 h, monolayers were rinsed once with 10 mM Tris (pH 8.0)–1 mM EDTA, harvested into 0.5 ml of the same buffer per 10-cm dish, and transferred to 1.5-ml Eppendorf tubes. Fifteen microliters of 20% (wt/vol) sodium dodecyl sulfate was then added, and the suspension was mixed gently until clear. This was left at room temperature for 15 to 20 min, after which 0.125 ml of 5 M NaCl was added; the samples were mixed and placed at 4°C overnight. Next, the precipitate was pelleted, and the supernatant was removed to a fresh tube and treated with proteinase K (50 µg/ml) at 37°C for 30 min, followed by one extraction with the phenol mixture noted above. The samples were precipitated with ethanol, dried, and digested in 0.1 ml of pancreatic RNase, 100 µg of 5 mM Tris-hydrochloride (pH 7.5) per ml–1 mM EDTA at 37°C for 30 min. Sodium chloride was added to 0.2 M, and the samples were again phenol extracted once and precipitated with ethanol. The dried pellets were dissolved in water, and samples were taken for appropriate restriction digestion.

DNA sequencing. The chemical degradation methods of Maxam and Gilbert (33) were followed for the

sequence determination of 5'-³²P-labeled cDNAs and of mutant virus constructs. In general, restricted, mutant viral DNAs were 3' labeled with ³²P at the unique *Msp*I site by limited repair synthesis with *E. coli* DNA polymerase, large fragment, and [α -³²P]dCTP. The restriction fragments *Bgl*I-*Msp*I (m.u., 0.661 through 0.725; residues 5157 through 265) and *Msp*I-*Sau*96I (m.u., 0.725 through 0.766; residues 265 through 478) were separated on a 5% polyacrylamide gel, recovered (as above), and subjected to sequencing methods.

RESULTS

As noted above, the analysis of various of the SV40 mutations such as deletions and substitutions within the templates for the late leader sequences showed several unexpected phenomena. Pursuing this line of investigation, we studied mRNA produced by two naturally occurring variants of SV40 and by several constructed mutants.

The primary method of analysis was to determine the sequence of cDNA products produced

FIG. 2. Construction scheme for *dlKMTD3/2-7*.

by reverse transcription of RNA. As noted previously by us, however, certain cautions should be considered in interpreting the data obtained (41). This method generally allows for the location of the 5' ends of RNA transcripts with an uncertainty of ± 2 to 3 residues due to sequencing pattern vagaries. In addition, although the locations of the principal 5' ends of wild-type SV40 RNA labeled in vivo have corresponded to those deduced by reverse transcription, the difficulty in obtaining suitable amounts of in vivo-labeled RNA has not allowed us to corroborate the assignments of the 5' ends of minor forms or very heterogeneous RNAs. It is possible that

some of the cDNAs result from premature termination of reverse transcripts or transcription of degraded RNAs. In selecting cDNA products for analysis, we have concentrated only on those that are most prominent or those that are representative of the "extension pattern" obtained (or both). We have disregarded very minor and background cDNA bands. In support of our 5' end assignments, similar deduced 5' ends for late RNA have appeared consistently in our analyses of many different viruses (14, 41; this report). Also, the existence and position of several of the less prominent 5' end sites have been confirmed by utilizing cDNA primers lo-

cated proximal to the proposed RNA termini (17, 41).

Naturally occurring variants with genome alterations upstream of the major RNA initiation site. 777N and SP1, two spontaneous variants of SV40 that have unusual alterations in the DNA template for the late leader, were available to us (37). The sequences of the pertinent portions of the two viruses, strain 777N and strain SP1, are shown in Fig. 3b and c. Both have a 17-base-pair (bp) deletion encompassing the *Kpn*I site (Fig. 3a); SP1 also contains a 47-bp substitution at that point. SP1 is apparently derived from strain 777N and has evolved a nearly perfect duplication of the DNA template encoding the major 5' end of late SV40 mRNA at residue 243. The sequence from residue 214 through 257 is dupli-

cated with three sequence alterations. Residues 228' and 252' to 253' upstream and downstream, respectively, of the duplicated 243' residue are deleted, and there is an insertion of 6 bp adjacent upstream to the duplicated 243' residue.

As determined by primer extension analyses (Fig. 4a and b, Fig. 5a) and confirmed by DNA sequencing of the cDNAs, the late mRNA produced by both of these strains (777N 19S RNA was not examined) showed several interesting deviations from the patterns of wild-type SV40 late mRNA leader sequences (data not shown; 17, 42). These results are summarized in Tables 1 and 2 and Fig. 7 and 8. First, the viral RNA from these variants contains large amounts of 16S leader sequence with tandem replication of the DNA sequence from residues 352 through

FIG. 3. Sequence structure of wild-type and mutant viruses. Shown are the sequences of the late sense strands for the relevant region of the genomes. Residue numbering of mutant DNAs is relative to that of the wild-type (strain 776) sequence. Tandem duplicated residues are written in lowercase and are designated by prime numbers. The *Kpn*I and *Msp*I restriction sites used in the construction of the mutant genomes and the major late RNA capping site (*) are indicated in the wild-type sequence (a). Symbols: [], sequence deleted between indicated nucleotides, the deleted sequence is shown above the brackets within parentheses; + + +, inserted nucleotides.

FIG. 4. Autoradiograms of 16S RNA extension products. The 5'-³²P-labeled DNA primer generated by *Hae*III and *Hind*III digestion and extending from residue 1382 to residue 1415 (0.939 to 0.945 m.u.) was hybridized to polyadenylated cytoplasmic late RNA and extended with reverse transcriptase. The cDNA products were separated on 8% polyacrylamide-7 M urea gels, 2 mm in thickness, in the Tris-borate-EDTA buffer system of Peacock and Dingman (40) and detected by autoradiography at 5°C. The numbered cDNAs were recovered from the gel and subjected to the chemical degradation sequencing method of Maxam and Gilbert (33). The results are given in Table 1. The specific extension patterns are for the late RNAs of (a) 777N, (b) SP1, (c) KMTD3, and (d) *dl*KMTD3/2-7.

444. This replication differs by one additional residue (a G) from that reported previously by this laboratory for a minor wild-type 16S RNA species (42) and is probably a correction of the latter. Thus, the replications are flanked by sequences corresponding to canonical splice junctions (47). For both viruses, there are clearly detectable leader segments with triple and possibly even quadruple tandem copies of this RNA sequence, although it is encoded only once in the viral DNA (Fig. 4a and b; Table 1). The abundance of these tandem repeats is intermediate between that previously observed for the wild type and the abundance of repeats de-

scribed for polyomavirus (32). As conclusively demonstrated for polyomavirus (32, 51) and inferred for SV40 (12, 42), these repeats are the result of splicing events that remove approximately a one-genome-length RNA fragment from a larger transcript. From rough estimates of autoradiograph band intensities, as much as 25% of 16S RNA of 777N and SP1 contains more than one copy of the repeated sequence (see Fig. 7).

Second, the 19S late mRNA of SP1 contains a form of leader that appeared to contain an internal deletion of five nucleotides (Fig. 5a; Table 2). Examination of the sequence suggests that this may arise by an internal splice removing an RNA copy consisting of one genome length plus five nucleotides of viral RNA. This splice apparently occurs from residue 239' in the

FIG. 5. Autoradiograms of 19S RNA extension products. The methods are as described in the legend to Fig. 4, except that the 5'-³²P-labeled DNA primer generated by *Sau*96I and extending from residue 479 to 509 (0.766 to 0.772 m.u.) was utilized. The results of sequencing analyses are listed in Table 2. The specific extension patterns are for the cytoplasmic late RNAs of (a) SP1 and (b) KMTD2 and for the nuclear late RNA of (c) KMTD3. In (c) only the portion of the autoradiogram depicting the extension products for the major RNA 5' end locations, 243' and 243, is shown for emphasis.

TABLE 1. 16S RNA (cDNA) species

Virus	cDNA product (no.)	Type of splice ^a	Approximate location of 5' terminus (residue)
777N	1	Continuous	?
	2	Continuous unspliced	850 ± 60?
	3	Continuous unspliced	900 ± 60?
	4	S	?
	5	S, TT	243
	6	S	130–140
	7	S, TD	243
	8	S	180–190
	9	S	~190
	10	S	243
	11	S	273
	12	S	283
	13	S	289
	14	S	310
	15	S	316
	16	S	342
	17	S	365
	18	S	383
SP1	1	Continuous unspliced	?
	2	Continuous unspliced	?
	3	Continuous unspliced	?
	4	S, T?	?
	5	S, TT	230?
	6	S, TT	243?
	7	S	?
	8	S, TD	230
	9	S, TD	243
	10	S	214–225
	11	S	230
	12	S	243
	13	S	260–270
	14	S	280–285
KMTD3	15	S	305–310
	16	S	314–318
	17	S	324
	18	S	342
	19	S	351–353
	20	S	366
	21	S	385
	22	S	401–403
	1	?	?
	2	S	182–187
	3	S	230'
	4	S	243'
	5	D	~87
	6	S	243
	7	D	~120
	8	D	132–137
	9	D	145–150
	10	D	157–160
	11	D	160–165
	12	D	176–180
	13	D	~182
	14	S	360

TABLE 1—Continued

Virus	cDNA product (no.)	Type of splice ^a	Approximate location of 5' terminus (residue)
d/KMTD3/2-7	1	Continuous unspliced	?
	2	Continuous unspliced	?
	3	S	182
	4	S	? (around deleted sequence)
	5	S	230'
	6	S	236'–238'
	7	S	243'
	8	S	243
	9	S	270–273
	10	S	289
	11	S	316
	12	S	366

^a S, Standard, single-spliced species, splice from residue 444 to 1381; TT, contains a tandem triplication of the sequence from residue 352 through 444; TD, contains a tandem duplication of the sequence from residue 352 through 444; D, double-spliced species, splices from residues 212 to 352 and 444 to 1381.

upstream copy of the late leader sequence to a residue 6 nucleotides downstream and follows the canonical splicing rules (47) so that an apparent deletion is produced. The ready detection of this in SP1 correlates with the large amount of tandemly replicated 16S mRNA leader. This is the only case in which we have noticed a splice in the leader sequence of SV40 mutants which is not already present in wild-type virus. It is noteworthy that the splice sites used are generated by the 6-bp insert in the duplicated leader template. No 19S RNA having the duplicated leader structure noted in 16S RNA was detected.

Third, although the major start for late RNA at 243 and 12 of 14 nucleotides immediately downstream are duplicated in SP1, the upstream, duplicate 243' site is utilized minimally and is found only in 19S RNA (see Fig. 7 and 8).

Fourth, a strong 5' start site at approximately residue 230, which is a T and lies in the middle of a string of 10 pyrimidines, is apparent in the late RNA of SP1. Recently, Cowie et al. (8) showed in polyomavirus infected cells that the first capped nucleotide or two of the 5' end of minor forms of early mRNA may be "borrowed" from transcripts of slightly upstream sequences, rather than initiated with a pyrimidine. Further investigation will be necessary, although it appears that this is an attractive alternative possibility for the structure of the 5' end of RNA located at residue 230.

TABLE 2. 19S RNA (cDNA) species

Virus	cDNA product (no.)	Type of splice ^a	Approximate location of 5' terminus (residue)
SP1	1	?	?
	2	?	?
	3	?	?
	4	U + S (mixture)	?
	5	U	243
	6	S	236-238
	7	S	243
	8	mostly L, X (some S)	?
	9	L, X	182
	10	L, X	190
	11	L	219'-
	12	L	225'
	13	L	243'
	14	L	230
	15	U	243
KMTD2	1	?	444
	2	?	?
	3	?	?
	4	?	?
	5	S	?
	6	S	?
	7	L	187
	8	L	?
	9	L	190-200
	10	L	208
	11	SL	87-92
	12	L	230'
	13	L	236'
	14	L	243
	15	L	?
	16	L	224
	17	SL	157
	18	L	243
	19	L	?
	20	SL	178
	21	SL	180

^a U, Unspliced species, lacking a splice upstream from residue 476; S, short splice, splice from residue 444 to 476; L, long splice, splice from residue 291 to 476; SL, super long splice, splice from residue 212 to 476; X, lacks residues 239' through 242' + 2, apparent splice.

Of the observations noted above, our attention focused on the abundance of 16S RNA leaders containing tandem replicated sequences. Earlier analysis of wild-type RNA (17, 42) indicated a correlation between the selection of particular 5' ends for 16S RNA and leader structure reflected either by an intraleader splice or a tandem duplication of the leader sequence in question. The sequence changes noted within the leader template region of 777N and SP1 seemed insufficient to account for a presumable

overproduction of greater-than-unit-genome length precursor RNA which could give rise to the observed abundance of tandemly repeated leader sequences. This prompted us to examine the genomic structure of these viruses relative to that of the more extensively characterized wild-type virus, 776. We found several restriction polymorphisms and sequence differences (data not shown) from a limited analysis and are currently pursuing these observations. However, the investigation reported here was continued with mutant viruses constructed from wild-type 776 to avoid any potential complications derived from comparing mRNA generated by more genetically diverse viruses.

Sequence alteration upstream of the major RNA initiation site. For wild-type virus there is an internal splice within the leader of some 16S RNAs such that residue 212 is spliced to residue 352. No 16S mRNA whose 5' end lay upstream of residue 208 is present that does not have this intraleader splice. Conversely, the intraleader splice has never been detected in 19S mRNA with 5' ends located at or upstream from residue 208, so that there is a functional coupling between the intraleader splice and the downstream 16S splice. Partial uncoupling of these two splices was observed in *dl1626* and *dl1659* (41) mutants that lacked, respectively, the acceptor end of the intraleader splice or 48 nucleotides from the 5' portion of the spliced-out segment approaching and apparently interfering with the donor splice site. It seemed possible that the effects on the abundance of 16S and 19S mRNA with different 5' ends are secondary to the interference with this splice.

To investigate this, we constructed a group of mutants in which the sequences around the convenient *KpnI* restriction site at residue 212 had been altered by small deletions removing the splice donor signal at this site. Six mutants were examined in which the deletions ranged from 4 to 21 nucleotides and were located between residues 202 to 224. The extent of the deletions in these viruses also afforded us the opportunity to assess the influence of sequences normally -25 to -30 nucleotides from the major late RNA initiation site at residue 243. The sequences of the leader template regions encompassing the deletions of four of these viruses are shown in Fig. 3d through g. The sequences of the remaining two viruses having intermediate-size deletions were not determined.

Comparative primer extension analysis of the RNA produced by viruses with these small deletions of the upstream end of the intraleader splice showed that the principal 5' end of 16S mRNA is unaltered in position (Fig. 6, lanes d through i). However, with the mutant viruses all of the minor forms of 16S mRNA with longer

FIG. 6. Autoradiogram of the 16S primer extension products on the late RNAs of wild-type and several mutant viruses. The general methods and the labeled DNA primer used are as described in the legend to Fig. 4. This gel was 0.3 mm in thickness, and autoradiography was at -70°C since sequencing of the cDNAs was not planned. The positions of the cDNAs corresponding to the major late RNA leader whose cap site is at residue 243 and to the KMTD-specified RNAs having 5' ends at about 230' and 243' are indicated. The extension patterns are for the late RNAs of (a) KMTD3, (b) KMTD2, (c) 777N, (d) *dIK-S2*, (e) *dIK-M6*, (f) *dIK-L6*, (g) *dIK-L5*, (h) *dIK-L4*, (i) *dIK-L3*, and (j) WT776.

leader sequences presumably contain the segment of the leader normally spliced out in wild-type virus (Fig. 6, lanes d through i; cDNAs migrating slower than the principal wild-type form). Shifts in the exact location of the 5' ends of minor forms of 16S RNA may have occurred, but were not further investigated. The results have several implications.

First, the intraleader splice is not an obligatory precursor to the formation of the downstream 16S mRNA splice. Instead, the converse is suggested. It might be necessary for the 16S RNA splice to occur before or in conjunction with the upstream leader splice to account for the absence of the upstream leader splice in 19S mRNAs.

Second, the major location of 16S RNA 5' ends is at residue 243 even in mutants with a deletion substituting nucleotides more than 19 nucleotides upstream from this site. Thus, the sequences located at approximately the position of the TATA box in conventional polymerase II promoters could be altered by various deletions without affecting the *in vivo* initiation at this site. The normal level of utilization of residue 243 and the relatively normal distribution of upstream leader 5' ends show that this intraleader splice does not affect the relative abundance of accumulated RNA with 5' ends upstream of 243.

Third, the results again confirm the conclusion that the initiation codon used for translation of the major viral capsid protein, VP1, functions effectively even if it is preceded by two or more initiation codons present in the late leader sequence (for a review, see reference 31a). This has previously seemed likely, but the minor forms of wild-type 16S RNA with internal splices in the leader removing the upstream initiation codons theoretically could encode VP1. The ability to bypass the upstream initiation codons is true also with *dI1626*, which lacks the acceptor site of the intraleader splice (41). Unlike *dI1626*, though, the present mutants retain all of the coding sequences of the late leader.

Tandem duplication of the region surrounding the major RNA initiation site. To investigate further which sequences are important in locating the 5' ends of late RNA, we constructed and studied mutants of SV40 virus in which there are direct tandem duplications of the sequences immediately surrounding the template for the major 5' end at residue 243. The construction of these mutants is described above and the sequence structures of the resulting mutants used in the present studies are shown in Fig. 3h through j.

These viruses mimic the situation in the spontaneous mutant SP1 in which the sequences about residue 243 are duplicated, but lack the additional base changes noted in SP1. Residues 218 through 264 in KMTD2 and residues 215 through 264 in KMTD3 are duplicated.

The mutant *dIKMTD3/2-7* was constructed with a deletion encroaching on the tandem duplicated sequence from the upstream *KpnI* site. Derived from KMTD3, it has effectively a shorter tandem duplication than KMTD3 or KMTD2 so that only the 20 nucleotides preceding the upstream 243' site are preserved.

For KMTD2 and KMTD3, both 19S and 16S mRNAs were examined by primer extension and in alternate instances, as indicated below, by sequencing of cDNA extension products. These viruses, which differ in structure only by three

FIG. 7. Summary of the locations and utilization of 5' ends of 16S RNA directed by wild-type and mutant viruses. The vertical axis represents the relative abundance of individual 5' ends in arbitrary units from 0 to 5. The horizontal axis represents the DNA residue numbers described in Fig. 3. The contributions to the individual 5' ends made by the various spliced RNA forms are shown. Symbols: , contiguous leader sequence, no splices or replications; , contains intraleader splice of residues 212 to 352; , contains tandem duplication of residues 352 through 444; , contains tandem triplication of residues 352 through 444.

nucleotides in the length of the duplicated sequence, direct almost identical late RNA profiles in which the three nucleotide sequence difference is reflected in the lengths of the longer

leaders (Fig. 6a and b; data for 19S RNA not shown, but patterns are virtually identical). Therefore, it was necessary only to sequence the cDNA extension products of one 16S RNA

FIG. 8. Summary of the locations and utilization of the 5' ends of 19S RNA directed by wild-type and mutant viruses, as in the legend to Fig. 7. Symbols: , contains type 1 splice from residue 291 to 476; , contains type 2 splice from residue 444 to 476; , contains type 3 splice from residue 212 to 476; , type 4, no splice (17).

pattern (KMTD3) and one 19S RNA pattern (KMTD2) (Fig. 4c and 5b; Tables 1 and 2). The results are summarized in Fig. 7 and 8. There are several important observations. Each of the duplicate, wild-type, major RNA start sites, 243' and 243, is utilized, although neither specifies the most abundant 5' end. Sites upstream of 243' are abundantly utilized, and the 5' end at about residue 87 is the most prominent overall. This is so even though the sequence of the entire late leader region of wild-type SV40 from the origin of replication through 22 nucleotides beyond the major transcription initiation site is unaltered in KMTD2 and KMTD3. This is consistent with

the effect on RNA start sites seen for *dl1626* (41), the mutant whose deletion extends downstream from residue 265, and for *SVL₁-r₁-29* (20), a defective virus in which the late region beginning at residue 262 is replaced with the exons for the rat preproinsulin I gene. It is noteworthy that one of the two most abundant RNA 5' termini for the latter virus was located at SV40 residue 85. These results again emphasize the importance of downstream sequences in specifying the selection of late RNA 5' ends. Gross alteration of the sequences approximately 22 nucleotides downstream of a major wild-type start site (in this case, 243') shifts the abundantly

utilized sites upstream. Conversely, alteration of all sequences upstream of residue 218 (25 nucleotides upstream of the major wild-type RNA start site, residue 243) as a result of the duplication did not substantially decrease the utilization of the 243 site. This result is consistent with the results for the *dIK* series of viruses noted above.

A strong RNA start site at residue 230' was also noted. This is reminiscent of that reported above for SP1 but differs, however, in that the upstream 230' residue is utilized here rather than the downstream 230 residue utilized in SP1.

Although the KMTD viruses specify considerably longer and structurally altered late leaders, as with late leader region deletion mutants (14, 21, 41), no internal splice was noted that had not already been described for wild-type SV40 late RNA. Additionally, the relative abundance of 5' ends at residues 230' and 243' demonstrates uniquely that large amounts of SV40 16S RNA with long leader sequences lacking internal splices may be produced. An overall comparison of the location of 19S and 16S 5' ends also provided another example of the preferential use of RNA initiated at certain 5' sites (residues 190 through 210 and 224) for 19S RNA splicing and of RNA beginning at other sites (residue 182) for 16S RNA splicing. This phenomenon was noted previously for late RNA specified by wild-type SV40 (17) and by certain deletion mutants (41), although the exact sites vary with the virus studied. We also analyzed nuclear RNA from KMTD3 with the primer specific for 19S RNA (Fig. 5c) and with a primer hybridizing nearer the major 5' ends of the RNA from residue 262 to 278 (an *HhaI-MspI* SV40 DNA fragment, data not shown). Primers hybridizing further downstream on the nuclear precursor (including that used for 16S RNA analysis) were not employed since the resulting long cDNAs would be difficult to resolve and analyze. The results showed that all RNA species detected are unspliced and retain 5' ends represented in the cytoplasmic RNA. Approximately equal amounts of KMTD3 RNA begin at residue 243' and 243 (Fig. 5c), consistent with the results for cytoplasmic RNA (Fig. 7 and 8). In Fig. 5c, the contribution of cDNAs corresponding to RNAs beginning at 236' through 238' (and some beginning adjacent upstream) to the broad, relatively intense cDNA band indicated as 243' was taken into account. These results suggest that, at least for the major RNA species, the relative amounts of cytoplasmic RNA with different 5' ends from KMTD3 infected cells are not a consequence of differential degradation concurrent with or after splicing.

Finally, 16S RNA produced by the mutant *dIKMTD3/2-7* was examined. In this case, the upstream 243' site and the downstream 243 site

are the major sites utilized for RNA 5' ends, even though the sequences further upstream than 18 nucleotides and 28 nucleotides, respectively, are altered. The present data confirm and extend the earlier conclusion that there is no critical sequence upstream of the 243 site that corresponds to the positional TATA box.

It should also be noted that the abundant forms of late RNA produced by *dIKMTD3/2-7* have 5' ends lying in the upstream duplicated sequence from 230' to 243' and adjacent immediately upstream to it. Relative to the 16S RNA results for KMTD3 (parent virus to *dIKMTD3/2-7*), it seems that the effect of the secondary deletion about the *KpnI* site is both to enhance the abundance of the upstream 243 site (243' over 243) and to suppress the utilization of sites further upstream (no sites upstream of residue 182 and decreased utilization of site 182). This is true although the sequence upstream and for over 20 nucleotides downstream was the same as that in wild-type virus.

In addition, since each of these KMTD series mutants contains the entire coding sequence for the normal agnogene product and produces substantial amounts of RNA whose structure corresponds to that of the major late 16S mRNA, the infected cells would synthesize the "agnoprotein" (26, 29) encoded in the leader. Therefore, the various effects on upstream initiation sites displaced away from residue 243, etc., are probably not due to inactivation of the agnogene product as might be argued for deletion mutants (1, 36). Finally, the relative abundance of RNA containing tandem replications of the 3' portion of the late leader sequence in each of these mutants is not substantially different from that for the wild-type virus, affirming our impression that the abundances of tandemly duplicated RNA noted in 777N and in SP1 was not simply a consequence of alterations of the templates for the 5' ends of the RNA.

DISCUSSION

The present results again emphasize the remarkable flexibility of SV40 virus with regard to alterations in the DNA sequence encoding the late leader and shed additional light on several aspects of interactions involved in the biogenesis of the complex pattern of SV40 late mRNA. The results address specifically the question of what determinants influence the selection of 5' start sites for SV40 late RNA.

It is clear that sequences 19 nucleotides or more upstream of the major initiation site at residue 243 are not necessary to position a major 5' end of RNA at that site. The results of the RNA analysis for the *dIK* series of viruses and for 777N in which 243 is the major 5'-end site specifically eliminate the possibility that a se-

quence located -25 to -30 nucleotides from the major late transcript either could be functionally analogous to the TATA or Hogness-Goldberg box noted in other eucaryotic promoters or could play a predominant role in determining the abundance of that transcript. Longer cDNA products of 16S RNA directed by the *d/K* series of viruses were observed, whereas a shorter cDNA product noted for the wild-type virus was absent. These differences are likely due to the inability of *d/K* RNA to form the intraleader splice, with the result that 16S RNA beginning upstream of 243 has correspondingly longer leaders, rather than being due to a more direct influence of the upstream alteration on the selection of 5' ends. In addition, the downstream 243 site in KMTD2 and KMTD3 is utilized abundantly, even though one effect of the tandem construction is an alteration of the sequences upstream more than 26 and 29 nucleotides, respectively. Finally, the deletion in *d/KMTD3/2-7* comes within 19 nucleotides of the duplicated and abundantly utilized 243' start site. In this last example, it should be noted that the 243' site is utilized more frequently than it is in the parent virus KMTD3. This effect may be the result of a combination of influences, as discussed below, rather than an example of a direct influence of an upstream sequence on a major start site.

On the other hand, the overall upstream shift in the selection of abundantly utilized 5' ends in KMTD2 and KMTD3 cannot be a compensation for the inability to form "normal" transcripts, e.g., those beginning at residue 243, as could be argued for *d/1626* or other deletion-substitution mutants (1, 36, 41). The KMTD viruses obviously retain and utilize abundantly the "wild-type" 202 nucleotide long late leader beginning at residue 243. Hence, the abundant utilization of sites upstream of residue 211 must result from the influence of local sequences. Recently, Brady et al. (3) examined total cellular RNA directed by a deletion mutant, pSVs-KB6, which lacks the identical residues that *d/K-M6* lacks (Fig. 3). They report that the effect of the deletion was to decrease severalfold the utilization of the 243 start site in vivo and to promote the utilization of five alternative start sites 35 to 152 nucleotides upstream of the 243 site. However, the authors did not distinguish cytoplasmic from nuclear RNA and did not evaluate the contribution of the intraleader splice to the S1 mapping patterns of wild-type RNA.

Qualitatively, specification of the major 5' start site at residue 243 is independent of sequences in contiguity more than 19 nucleotides upstream or 21 nucleotides downstream. The duplicated 243' site is the major start site of 16S RNA for *d/KMTD3/2-7* and is a prominent start site for late RNA of KMTD2 and KMTD3.

Piatak et al. (41) previously reported that the 243 site was utilized, although to a lesser degree, for 16S and 19S late RNA specified by *d/1626*, even though the deletion in that virus begins 24 nucleotides downstream of that site. The results of our analysis on the late RNA of SP1 also support this finding and lead us to suggest that the sequences immediately adjacent to the major 243 site are most critical in specifying it as an RNA 5' site. In this instance, the 243' site of SP1 is not utilized in 16S RNA and is a minor start site for 19S RNA. The duplicated sequence in SP1 differs primarily from those of KMTD2 and KMTD3 in having a 6-bp insertion adjacent upstream to the duplicated 243' site. This argues against the simplest form of a "primerase" model, in which initiation begins wherever the first few nucleotides of the transcript match a preferred sequence. To the contrary, the results suggest that the first six nucleotides upstream of a 243 site are a component of the "recognition" sequence specifying initiation at 243.

Quantitatively, major nucleotide substitutions altering the sequences downstream of residue 264 have a significant effect on the frequency of expression of the upstream 243 site in late RNA. In KMTD constructs, there is a gross alteration of the sequence downstream of residue 264', and there is an apparent upstream shift in the selection of abundant 5' ends relative to 243'. This is analogous to the results observed previously for *d/1626* (41) which has only one copy of the 243 site and lacks residues 267 through 419. KMTD2 and KMTD3 also differ from *d/1626* in that they contain the entire agnogene-coding region and produce substantial amounts of RNA that is identical to wild-type mRNA for the agnoprotein product. This excludes the possibility that deletion of the agnogene was responsible for the shift in the location of 5' ends of SV40 late RNA. It thus indicates that the principal effects of the lesion in *d/1626* and perhaps of the substitution of insulin-coding sequences for SV40 late leader segments (20) are the direct result of the alteration of sequences downstream from 265 that influence the local upstream transcription initiation site.

It should be noted, however, that the decreased utilization of the upstream 243' site in KMTD2 and KMTD3 is not as pronounced as that observed for *d/1626*. In fact, the results of the analysis on 16S RNA of *d/KMTD3/2-7* contrast with those of the KMTD3 parent in that the upstream 243' site is utilized more than the 243 site.

Clearly, the mechanism for determining and selecting SV40 late RNA start sites is unconventional relative to the current model for polymerase II transcription units. Some correlation can be made between gross sequence alterations in

180 220 243

ACATTCCACACT TTAGTTTC CC6GTT

and in this report (Fig. 5c), the patterns of 5'-end locations for extracted nuclear RNAs approximate that of the stable cytoplasmic species. Also, in the presence of actinomycin D, minor cytoplasmic RNA species with 5' ends other than the major forms were as stable as the major RNA species (41). The possibility that 5' ends are generated by the processing of a longer transcript or that the same (or other) hairpin-like structures are recognized by processing enzymes which generate such ends (or both) seems unlikely. Contreras and Fiers (6) established previously that the principal SV40 late RNA capped ends correspond to transcription initiation sites. Also, to date only vaccinia virus-infected cells are known to have the capacity to add caps to RNA with 5'-terminal monophosphates (48). Although it is possible that such a model for the selection of SV40 late RNA 5' ends might function, it also seems insufficient to account for the extreme diversity in the location and abundance of late RNA 5' ends in the wild-type and leader region mutant virus systems. It is more likely that selective transcript conservation is a secondary feature of the accepted model of transcription initiation in which the RNA polymerase recognizes certain sequences and initiates at specific sites in the DNA template. In this regard, the DNA equivalent of the potential, strong hairpin-like structures described for the 5'-terminal RNA sequence for the SV40 late region can be considered. The possibility must be seriously entertained that the DNA at the time of transcription initiation has melted in such a way that a particular downstream secondary structure is formed that, together with local sequences at the start site, specifies the site of initiation. The secondary structures hypothesized for the wild-type late leader template are

shown in Fig. 9. These are analogous to the structures drawn in the RNA sequences (21) and discussed above and that have been hypothesized to play a role in transcript attenuation (24). A multifunctional potential for analogous structural features is intriguing.

Presumably, the advancing DNA polymerase in a replicating fork or the interaction of RNA polymerase with the template (or both) could help generate such structures. Groudine and Weintraub (19) have speculated that similar structures are formed at globin DNase I-hyper-sensitive sites and are stabilized by the binding of a protein to single-stranded DNA. In the SV40 DNA sequence, the potential of a strong hairpin loop immediately downstream from residue 243 has been described (16) and observed as a contribution to the distortion in DNA sequencing gels of this region (52). Possibly related to this, Saragosti et al. (44) reported that in naked SV40 DNA, in contrast to SV40 chromatin, there is a very sensitive DNase I cleavage site at or just downstream of this region.

The hairpin loop upstream of site 243 (Fig. 9) cannot be strictly necessary for initiation at residue 243 since the *dIK* viruses and 777N cannot form this loop, yet 243 is utilized in each as the major initiation site. The KMTD viruses present a special, unique possibility. The upstream loop, between sites 182 and 243', can be formed as well as the loop downstream of the 243 site. No hairpin loop or secondary structure of similar stability can exist between the two copies of the 243 site. Several observations relative to the placement of abundant RNA 5' ends can be made. First, as in the *dIK* and 777N viruses, there is no potential loop preceding the downstream 243 initiation site in KMTD2, KMTD3, and *dIKMTD3/2-7*, yet this is a prominent RNA start site for each virus. Second, the upstream hairpin is intact in KMTD2 and KMTD3, but missing from the derivative *dIKMTD3/2-7*, consistent with the abundant use of the 182 site and further upstream sites by the former viruses but not by the latter virus. It is noteworthy that RNA with 5' ends at or upstream of 182 is absent with the variant SP1, which also lacks the upstream loop due to a deletion of residues 194 to 213. The preferential location of RNA 5' ends at or just upstream of 243' in *dIKMTD3/2-7* is especially revealing when considered in context with the effects produced by the lesion in its parent virus, KMTD3. It would appear that displacement of a hairpin loop downstream of 243' decreases utilization of that site, but that the additional alteration of the secondary structure upstream of 243' partially restores that utilization.

Results from studies of the late RNA produced by other deletion mutant viruses (14, 21,

22, 41, 55) are not inconsistent with our model, although many are difficult to evaluate. In most instances, an extensive portion of the late leader template is removed, including sequences encoding major wild-type RNA 5' ends and sequences involved in both hairpin structures shown in Fig. 9. In general, the major 5' ends of the mutant RNAs are heterogeneous, lie upstream of the deletion, and are either absent or less abundant in wild-type RNA. In some instances, abundant RNA 5' ends are located downstream of the deletion (14, 41). It is also consistent with the above model and the results for these deletion mutants that elimination of the entire stem and loop region shown in Fig. 9 allows for less ordered selection of 5' termini in the possible open template, nucleosome-void region of the viral chromosome.

We have also made observations on the correlation of specific late RNA 5' ends with splicing events and on the regulation of splicing events in SV40 late RNA. In some cases, there is an apparent selection of certain 5' ends for 19S RNA but not for 16S RNA, e.g., the 5' ends at residues 190 through 210 in RNA produced by the KMTD viruses, reminiscent of our observations on wild-type RNA (17). In that instance, late RNA with 5' ends upstream of residue 243 contribute relatively more to 19S than to 16S mRNA, as compared with RNA initiated at 243. The length of the leader sequence is certainly not the sole or principal determinant of the distribution since the KMTD viruses produce large amounts of 16S RNA beginning at the upstream 230' and 243' residues and lacking any intraleader splice. The 16S leaders in this case are at least 50 nucleotides longer than the major wild-type 16S leader.

We have considered several possible explanations for the selection existing in 19S versus 16S RNA. First, it may be due to an artifact of analysis in that a 16S RNA may be made with a 5' end at residues 190 through 210 and containing the intraleader splice from residue 212 to 352. A cDNA copy of such a transcript would have a short sequence beyond the acceptor site at 352, and we have noted a 16S cDNA of approximately the appropriate length (Fig. 4c, no. 14). We have confirmed by sequence analysis, however, that this cDNA does stop short of 352 at nucleotide 360. Hence, this artifact of analysis does not seem likely. Second, it may be argued that 16S transcripts beginning at residues 190 through 210 are inherently unstable and do not appear in the cytoplasm. We addressed this argument for the wild-type virus previously (41) and for KMTD3 in this report and observed each time that the distribution of 5' ends of unspliced nuclear precursors was similar to that for the spliced cytoplasmic forms. The relative amounts

of nuclear RNA with 5' ends at residues 190 through 210 are not different from that found for the cytoplasmic species starting at residues 190 through 210. Hence, a possible unstable transcript is unlikely unless it were turned over very rapidly in the nucleus. Third, the selection of a particular 5' end could directly determine the downstream splice formed independent of any effect of downstream sequences. Our results on other mutant viruses weigh against this. Piatak et al. (41) demonstrated for the mutants *dl1659* and *dl1626* that any 5' end could exist for either downstream 16S or 19S splice. Fourth, the structure of the leader per se would determine which downstream splice is formed. Hence, the selection of a 5' end indirectly influences the downstream splicing event in that a particular leader sequence would result. This agrees with and extends the conclusions of Ghosh et al. (14) in explaining the aberrant ratios of 16S to 19S RNA noted for different mutants in which the major 5' end sites and large portions of the common wild-type leader template are deleted. Thus, this final argument remains the most plausible to us and does not conflict with our argument that the downstream 16S splice may be a prerequisite for the formation of leader splices.

The ability of the KMTD viruses to form 16S mRNA with long leader sequences lacking the intraleader splice (residues 212 to 352) suggests that the removal of downstream sequences or execution of the downstream splice may be a necessary prerequisite, directly or indirectly, for the formation of the upstream leader splice, rather than the other way around, although more complicated explanations cannot be excluded. Further, although the internal splice in the leader of SV40 16S mRNA removes any initiation codons preceding the VP1 initiation codon from forms of 16S mRNA, such an alteration is clearly not necessary for the ribosomes to be able to bypass the two initiation codons in the leader and to start at the initiation codon for VP1.

The results of the analyses on late RNA specified by SP1 indicate that the correlation between the 16S leader to body splice and the intraleader splice mentioned above is specifically a correlation between the 16S leader to body splice and the utilization of the splice acceptor at residue 352. Late 16S RNA containing tandem repetitions of the 3' portion of the leader from residue 352 through 444 is prominent (Fig. 7). Such tandem structures presumably are the result of a splice from the donor site at residue 444 to the acceptor site at residue 352 in a greater-than-genome-length precursor (42). Although the same long precursor would be available for 19S RNA splicing, no splice utilizing the acceptor site at residue 352 was detected in 19S RNA of SP1 (Table 2). A similar observation had been

noted previously for the wild-type system (16, 42), but the very low occurrence of a tandem leader sequence in wild-type 16S RNA made it impossible to exclude its occurrence in 19S RNA.

In conclusion, the positioning of SV40 late RNAs is clearly the result of DNA sequence elements quite different from the TATA box of the "canonical" polymerase II transcription units that have been studied most extensively. Sequences proximal (likely within 20 nucleotides) to a major RNA start site seem to be most important in selecting a site, although the level to which that site is utilized is apparently modulated by more distal downstream sequences (greater than about 21 nucleotides). The influence of transient, alternative, local DNA structures on both the selection and modulation of RNA start sites should also be considered. In view of the small amount of SV40 genetic information and the several functions already assigned to the SV40 early proteins, it is attractive to consider that some subset of host genes may have their transcription under similar types of control.

ACKNOWLEDGMENTS

We thank Donna Mitchell and Sharlene Ivory for excellent technical assistance and Marjorie Veronneau and Kathy Levenson for their help in preparing this manuscript.

This work was supported by a grant from the American Cancer Society and by Public Health Service grants RO1 CA32799 and CA16038 awarded by the National Cancer Institute.

LITERATURE CITED

1. Alwine, J. C. 1982. Evidence for simian virus 40 late transcriptional control: mixed infections of wild-type simian virus 40 and a late leader deletion mutant exhibit trans effects on late viral RNA synthesis. *J. Virol.* 42:798-803.
2. Benoist, C., and P. Chambon. 1981. *In vivo* sequence requirements of the SV40 early promoter region. *Nature (London)* 290:304-310.
3. Brady, J., M. Radonovich, M. Vodkin, V. Natarajan, M. Thoren, G. Das, J. Janki, and N. P. Salzman. 1982. Site-specific base substitution and deletion mutations that enhance or suppress transcription of the SV40 major late RNA. *Cell* 31:625-633.
4. Canaani, D., C. Kahana, A. Mukamel, and Y. Groner. 1979. *Proc. Natl. Acad. Sci. U.S.A.* 75:3078-3082.
5. Chu, G., and P. A. Sharp. 1981. SV40 DNA transfection of cells in suspension: analysis of the efficiency of transcription and translation of T-antigen. *Gene* 13:197-202.
6. Contreras, R., and W. Fiers. 1981. Initiation of transcription by RNA polymerase II in permeable, SV40-infected or noninfected, CV1 cells; evidence for multiple promoters of SV40 late transcripts. *Nucleic Acids Res.* 9:215-236.
7. Corden, J., B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedlinger, and P. Chambon. 1980. Promoter sequences of eukaryotic protein-coding genes. *Science* 209:1406-1414.
8. Cowie, A., P. Jat, and R. Kamen. 1982. Determination of sequences at the capped 5' ends of polyoma virus early region transcripts synthesized *in vivo* and *in vitro* demonstrates an unusual microheterogeneity. *J. Mol. Biol.* 159:225-255.
9. Cowie, A., C. Tyndall, and R. Kamen. 1981. Sequences at

- the capped 5' ends of polyoma virus late region mRNAs: an example of extreme terminal heterogeneity. *Nucleic Acids Res.* 9:6305-6322.
10. Davies, R. L., S. Fuhrer-Krusi, and R. S. Kucherlapati. 1982. Modulation of transfected gene expression mediated by changes in chromatin structure. *Cell* 31:521-529.
 11. Dierks, P., A. Van Ooyen, N. Mantei, and C. Weissman. 1981. DNA sequences preceding the rabbit β -globin gene are required for formation in mouse L cells of β -globin RNA with the correct 5' terminus. *Proc. Natl. Acad. Sci. U.S.A.* 78:1411-1415.
 12. Ford, J. P., and M.-T. Hsu. 1978. Transcription pattern of *in vivo*-labeled late simian virus 40 RNA: equimolar transcription beyond the mRNA 3' terminus. *J. Virol.* 28:795-801.
 13. Ghosh, P. K., P. Lebowitz, R. J. Frisque, and Y. Gluzman. 1981. Identification of a promoter component involved in positioning the 5' termini of simian virus 40 early mRNAs. *Proc. Natl. Acad. Sci. U.S.A.* 78:100-104.
 14. Ghosh, P. K., M. Piatak, J. E. Mertz, S. M. Weissman, and P. Lebowitz. 1982. Altered utilization of splice sites and 5' termini in late RNAs produced by leader region mutants of simian virus 40. *J. Virol.* 44:610-624.
 15. Ghosh, P. K., V. B. Reddy, M. Piatak, P. Lebowitz, and S. M. Weissman. 1980. Determination of RNA sequences by primer directed synthesis and sequencing of their cDNA transcripts. *Methods Enzymol.* 65:580-595.
 16. Ghosh, P. K., V. B. Reddy, J. Swinscoe, P. V. Choudary, P. Lebowitz, and S. M. Weissman. 1978. The 5'-terminal leader sequence of late 16S mRNA from cells infected with simian virus 40. *J. Biol. Chem.* 253:3643-3647.
 17. Ghosh, P. K., V. B. Reddy, J. S. Swinscoe, P. V. Choudary, P. Lebowitz, and S. M. Weissman. 1978. Heterogeneity and 5' terminal structures of the late RNAs of simian virus 40. *J. Mol. Biol.* 126:813-846.
 18. Grosschedl, R., and M. Birnstiel. 1980. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.* 77:1432-1436.
 19. Groudine, M., and H. Weintraub. 1982. Propagation of globin DNaseI-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. *Cell* 30:131-139.
 20. Gruss, P. A., E. Efstratiadis, S. Karathanasis, M. Konig, and G. Khoury. 1981. Synthesis of stable unspliced mRNA from an intronless simian virus 40-rat preproinsulin gene recombinant. *Proc. Natl. Acad. Sci. U.S.A.* 78:6091-6095.
 21. Haegeman, G., D. Iserentant, D. Gheysen, and W. Fiers. 1979. Characterization of the major altered leader sequence of late mRNA induced by SV40 deletion mutant dl-1811. *Nucleic Acids Res.* 7:1799-1814.
 22. Haegeman, G., H. Van Heuverswyn, D. Gheysen, and W. Fiers. 1979. Heterogeneity of the 5' terminus of late mRNA induced by a viable simian virus 40 deletion mutant. *J. Virol.* 31:484-493.
 23. Hashimoto, S., and M. Green. 1980. Adenovirus 2 early messenger RNA-genome mapping of 5'-terminal RNase T1 oligonucleotides and heterogeneity of 5'-termini. *J. Biol. Chem.* 255:6780-6788.
 24. Hay, N., H. Skolnik-David, and Y. Aloni. 1982. Attenuation in the control of SV40 gene expression. *Cell* 29:183-193.
 25. Herbolmel, P., S. Saragosti, D. Blangy, and M. Yaniv. 1981. Fine structure of the origin-proximal DNase I-hypersensitive region in wild-type and EC mutant polyoma. *Cell* 25:651-658.
 26. Jackson, V., and R. Chalkley. 1981. Use of whole-cell fixation to visualize replicating and maturing simian virus 40: identification of new viral gene product. *Proc. Natl. Acad. Sci. U.S.A.* 78:6081-6085.
 27. Jakobovits, E. B., S. Bratosin, and Y. Aloni. 1980. A nucleosome-free region in SV40 minichromosomes. *Nature (London)* 285:263-265.
 28. Jakobovits, E. B., S. Saragosti, M. Yaniv, and Y. Aloni. 1980. *Escherichia coli* RNA polymerase *in vitro* mimics simian virus 40 *in vivo* transcription when the template is viral nucleoprotein. *Proc. Natl. Acad. Sci. U.S.A.* 77:6556-6560.
 29. Jay, G., S. Nomura, C. Anderson, and G. Khoury. 1981. Identification of the SV40 agnogene product: a DNA binding protein. *Nature (London)* 291:346-349.
 30. Keene, M. A., V. Corces, K. Lowenhaupt, and S. C. R. Elgin. 1981. DNase I hypersensitive sites in *Drosophila* chromatin occur at the 5' ends of regions of transcription. *Proc. Natl. Acad. Sci. U.S.A.* 78:143-146.
 31. Larsen, A., and H. Weintraub. 1982. An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. *Cell* 29:609-622.
 - 31a. Lebowitz, P., and S. M. Weissman. 1979. Organization and transcription of the simian virus 40 genome. *Curr. Top. Microbiol. Immunol.* 87:43-172.
 32. Legon, S., A. Flannell, A. Cowie, and R. Kamen. 1979. Amplification in the leader sequence of "late" polyoma virus mRNAs. *Cell* 16:373-388.
 33. Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* 65:499-560.
 34. McGhee, J. D., W. I. Wood, M. Dolan, J. D. Engel, and G. Felsenfeld. 1981. A 200 base pair region at the 5' end of the chicken adult β -globin gene is accessible to nuclease digestion. *Cell* 27:45-55.
 35. McKnight, S. L., E. R. Gravis, R. Kingsbury, and R. Axel. 1981. Analysis of transcriptional regulatory signals of the HSV thymidine kinase gene: identification of an upstream control region. *Cell* 25:385-398.
 36. Mertz, J. E., A. Murphy, and A. Barkan. 1983. Mutants deleted in the agnogene of simian virus 40 define a new complementation group. *J. Virol.* 45:36-46.
 37. Norkin, L. C. 1979. Small plaque variants of simian virus 40 from a persistent infection of rhesus monkey kidney cells. *Virology* 97:201-206.
 38. Norkin, L. C., and M. Piatak. 1982. Recombinant joints in a simian virus 40 variant generated in a persistent infection. *J. Gen. Virol.* 63:517-522.
 39. Osborne, T. F., R. B. Gayer, and A. J. Berk. 1982. The TATA homology and the mRNA 5' untranslated sequence are not required for expression of essential adenovirus E1A functions. *Cell* 29:139-148.
 40. Peacock, A. C., and C. W. Dingman. 1968. Molecular weight estimation and separation of ribonucleic acid by electrophoresis in agarose-acrylamide composite gels. *Biochemistry* 7:668-674.
 41. Piatak, M., K. N. Subramanian, P. Roy, and S. M. Weissman. 1981. Late messenger RNA production by viable simian virus 40 mutants with deletions in the leader region. *J. Mol. Biol.* 153:589-618.
 42. Reddy, V. B., P. K. Ghosh, P. Lebowitz, and S. M. Weissman. 1978. Gaps and duplicated sequences in the leaders of SV40 16S RNA. *Nucleic Acids Res.* 5:4195-4214.
 43. Samal, B., A. Worcel, C. Louis, and P. Schedl. 1981. Chromatin structure of the histone genes of *D. melanogaster*. *Cell* 23:401-409.
 44. Saragosti, S., S. Cereghini, and M. Yaniv. 1982. Fine structure of the regulatory region of simian virus 40 minichromosomes revealed by DNase I digestion. *J. Mol. Biol.* 160:133-146.
 45. Saragosti, S., G. Moyne, and M. Yaniv. 1980. Absence of nucleosomes in a fraction of SV40 chromatin between the origin of replication and the region coding for the late leader RNA. *Cell* 20:65-73.
 46. Scott, W. A., and D. J. Wigmore. 1978. Sites in simian virus 40 chromatin which are preferentially cleaved by endonucleases. *Cell* 15:1511-1518.
 47. Seif, I., G. Khoury, and R. Dhar. 1979. BKV splice sequences based on analysis of preferred donor and acceptor sites. *Nucleic Acids Res.* 6:3387-3398.
 48. Spencer, E., D. Loring, J. Hurwitz, and G. Monroy. 1978. Enzymatic conversion of 5'-phosphate-terminated RNA

- to 5'-di- and triphosphate-terminated RNA. *Proc. Natl. Acad. Sci. U.S.A.* **75**:4793-4797.
49. Stalder, J., A. Larsen, J. D. Engel, M. Dolan, M. Groudine, and H. Weintraub. 1980. Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNase I. *Cell* **20**:451-460.
50. Sweet, R. W., M. V. Chao, and R. Axel. 1982. The structure of the thymidine kinase gene promoter: nuclease hypersensitivity correlates with expression. *Cell* **31**:347-353.
51. Treisman, R., and R. Kamen. 1981. Structure of polyoma virus late nuclear RNA. *J. Mol. Biol.* **148**:273-301.
52. Yan Heuverswyn, H., and W. Fiers. 1979. Nucleotide sequence of the *Hind*-C fragment of simian virus 40 DNA. *Eur. J. Biochem.* **100**:51-60.
53. Varshavsky, A. J., O. Sundin, and M. J. Bohn. 1978. SV40 viral minichromosomes: preferential exposure of the origin of replication as probed by restriction endonucleases. *Nucleic Acids Res.* **5**:3469-3477.
54. Varshavsky, A. J., O. Sundin, and M. Bohn. 1979. A stretch of "late" SV40 viral DNA about 400 bp long which includes the origin of replication is specifically exposed in SV40 minichromosomes. *Cell* **16**:453-466.
55. Villarreal, L. P., R. T. White, and P. Berg. 1979. Mutational alterations within the simian virus 40 leader segment generate altered 16S and 19S mRNAs. *J. Virol.* **29**:209-219.
56. Waldeck, K. W., B. Fohring, K. Chowdury, P. Gruss, and P. Sauer. 1978. Origin of DNA replication in papovavirus chromatin is recognized by endogenous endonuclease. *Proc. Natl. Acad. Sci. U.S.A.* **75**:5964-5968.
57. Wu, C. 1980. The 5' ends of *Drosophila* heat-shock genes in chromatin are sensitive to DNase I. *Nature (London)* **286**:854-860.
58. Wu, C., P. M. Bingham, K. J. Livak, R. Holmgren, and S. C. R. Elgin. 1979. The chromatin structure of specific genes. I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**:797-806.