MOLECULAR AND CELLULAR BIOLOGY, Dec. 1992, p. 5386-5393 Vol. 12, No. 12
0270-7306/92/125386-08$02.00/0

Definition of the Upstream Efficiency Element of the Simian Virus 40 Late Polyadenylation Signal by Using In Vitro Analyses

NANCY SCHEK, CHARLES COOKE, AND JAMES C. ALWINE*

Department of Microbiology and Molecular Biology Graduate Group, School of Medicine, 560 Clinical Research Building, 422 Curie Boulevard, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6142

Received 18 May 1992/Returned for modification 23 July 1992/Accepted 2 September 1992

The polyadenylation signal for the late mRNAs of simian virus 40 is known to have sequence elements located both upstream and downstream of the AAUAAA which affect efficiency of utilization of the signal. The upstream efficiency element has been previously characterized by using deletion mutations and transfection analyses. Those studies suggested that the upstream element lies between 13 and 48 nucleotides upstream of the AAUAAA. We have utilized in vitro cleavage and polyadenylation reactions to further define the upstream element. ^{32}P-labeled substrate RNAs were prepared by in vitro transcription from wild-type templates as well as from mutant templates having deletions and linker substitutions in the upstream region. Analysis of these substrates defined the upstream region as sequences between 13 and 51 nucleotides upstream of the AAUAAA, in good agreement with the in vivo results. Within this region, three core elements with the consensus sequence AUUUGURA were identified and were specifically mutated by linker substitution. These core elements were found to contain the active components of the upstream efficiency element. Using substrates with both single and double linker substitution mutations of core elements, we observed that the core elements function in a distance-dependent manner. In mutants containing only one core element, the effect on efficiency increases as the distance between the element and the AAUAAA decreases. In addition, when core elements are present in multiple copies, the effect is additive. The core element consensus sequence, which bears homology to the Sm protein complex-binding site in human U1 RNA, is also found within the upstream elements of the ground squirrel hepatitis B and cauliflower mosaic virus polyadenylation signals (R. Russnak, Nucleic Acids Res. 19:6449–6456, 1991; H. Sanfacon, P. Brodmann, and T. Hohn, Genes Dev. 5:141–149, 1991), suggesting functional conservation of this element between mammals and plants.

The polyadenylation signal for the simian virus 40 (SV40) late genes (Fig. 1) has been utilized extensively as a model system for analysis of the molecular mechanisms of the polyadenylation reaction. Besides the consensus sequence AAUAAA and the cleavage site 13 nucleotides downstream, sequence elements which affect efficiency of utilization of the polyadenylation signal in vivo have been identified both upstream and downstream of the AAUAAA (7, 11, 30, 31). In addition, the presence of a 3′ splice site has been shown to affect the efficiency of polyadenylation in vivo (9).

Two downstream elements (DSEs) have been identified by deletion mutagenesis and analysis in different systems. One element lies between 19 and 40 nucleotides downstream of the AAUAAA (DSE19/40) and was defined as the major element in the oocyte system (11). The other lies between 59 and 67 nucleotides downstream of the AAUAAA (DSE59/67) and was defined as the major element in the African green monkey kidney cell line CV-1, while DSE19/40 appeared to have only a moderate effect in these cells (31). Both of these elements are U rich. DSE59/67 has been shown to be a heterogeneous nuclear ribonucleoprotein C binding site and can function as a DSE with an AAUAAA in vitro (34). DSEs have been detected in many polyadenylation signals and have been shown to affect polyadenylation efficiency not only in vivo (4, 5, 10, 11, 14, 16, 21–23, 30, 31) but also in vitro (17, 29, 37). DSEs do not conform to a consensus sequence but fall into groups which are most easily described as either GU or U rich. The functions of DSEs may be subject to spatial constraints, since insertions which increase the distance between the AAUAAA and a DSE often decrease polyadenylation efficiency (15, 18, 21). Additionally, human T-cell leukemia virus type 1, which has more than 250 nucleotides between the AAUAAA and the DSE, has an RNA secondary structure that juxtaposes the two elements (1, 3).

In more recent in vivo studies, additional efficiency elements lying upstream of the AAUAAA (upstream elements [USEs]) have been defined in a variety of viral poly(A) signals, including SV40 late (7), ground squirrel hepatitis virus (27, 28), adenovirus type 2 (12), cauliflower mosaic virus (CaMV) (19, 32), and human immunodeficiency virus (HIV) type 1 (6, 8, 13, 33). Nonhomologous USEs from different poly(A) signals are functionally similar; for example, the USE from the HIV type 1 poly(A) signal can replace the USEs of the SV40 late (33) and ground squirrel hepatitis virus (28) poly(A) signals. Further, the existence of homology between some USEs and DSEs suggests functional similarity (33).

In previous in vivo studies of the SV40 late polyadenylation signal, we found that a USE(s) existed within 13 to 48 nucleotides upstream of the AAUAAA (7). In the present study, we have utilized HeLa cell nuclear extracts (24–26) to define the SV40 USE by in vitro cleavage and polyadenylation reactions. We find that the element consists of three core elements having the sequence AUUUGURA. Linker substitutions of these elements, both singly and in pairs, show that they function additively and in a distance-depen-

* Corresponding author.

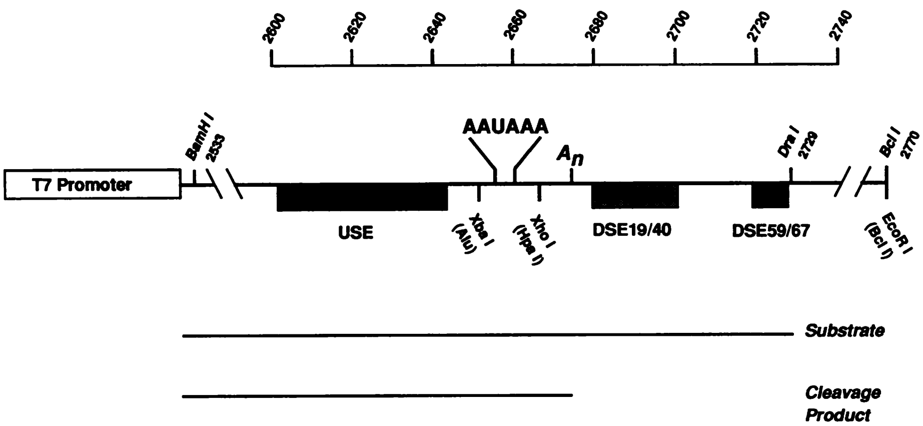
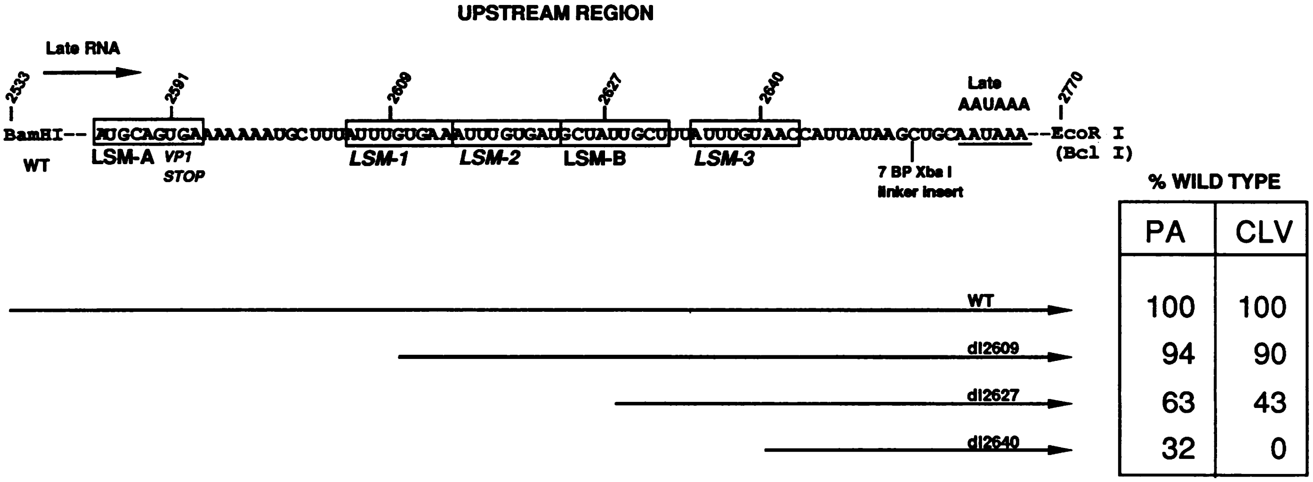


FIG. 1. The polyadenylation signal for the SV40 late mRNAs. SV40 sequences between the *Bam*HI and *Bcl*I sites (SV40 nucleotides 2533 to 2770) are shown as they appear inserted after the T7 promoter in pGEM2-PAS. These sequences contain all of the known polyadenylation signal elements including the AAUAAA as well as downstream efficiency elements DSE19/40 and DSE59/67 (11, 31) (see text) and upstream efficiency elements (USEs) (7). The sequence differs from the true SV40 sequence by the insertion of *Xba*I and *Xho*I linkers on opposite sides of the AAUAAA (7). These sites aid in the insertion and removal of USEs and DSEs. Polyadenylation using the signal with the linkers has been shown to be equivalent to the true wild-type signal in vivo (7). For in vitro transcription of polyadenylation substrate RNAs using T7 polymerase, the plasmids were linearized at the *Dra*I site (SV40 nucleotide 2729). Diagrams of the wild-type substrate and polyadenylation cleavage product are shown at the bottom. The exact sizes of substrates and cleavage products varied on the basis of whether the template was the wild type or a deletion mutant. Sizes of substrates and expected cleavage products are given in Table 1.

dent manner relative to the AAUAAA. The core elements we have defined are very similar to the defined USEs within the ground squirrel hepatitis virus (27, 28) and CaMV polyadenylation signals (19, 32), suggesting that the core elements provide a significant, evolutionarily conserved function in RNA processing. In addition, the core elements bear homology to the Sm protein complex-binding site on human U1 RNA (20).

MATERIALS AND METHODS

Plasmid templates for polyadenylation substrates. Substrate RNAs for in vitro polyadenylation reactions were prepared by in vitro transcription from plasmid templates containing the entire wild-type SV40 late polyadenylation signal or from templates having deletion or linker substitution mutations within the upstream region of this polyadenylation signal. The plasmid bearing the wild-type signal, pGEM2-PAS, contains SV40 nucleotides 2533 to 2770 inserted between the *Bam*HI and *Eco*RI sites of the in vitro transcription vector pGEM2 (Promega Biotec) (7). As indicated in Fig. 1 and 2, this sequence differs from the true SV40 wild-type sequence by the insertion of an *Xba*I site at SV40 nucleotide 2652 (formerly an *Alu*I site) and the insertion of an *Xho*I site at nucleotide 2668 (formerly a *Hpa*I site). These sites were previously inserted to aid in the removal and insertion of



% WILD TYPE

	PA	CLV
WT	100	100
dl2609	94	90
dl2627	63	43
dl2640	32	0

FIG. 2. The sequence of the upstream region of the SV40 late polyadenylation sequence. The sequences shown represent the significant region defined by deletion analysis in vivo (7). Boxed elements LSM-1, -2, and -3 are discussed in the text and represent the sequences specifically replaced with linkers. The sequences deleted in dl2609, dl2627, and dl2640 are indicated by the diagrams of the substrate RNAs shown below the sequence. The table on the right shows the quantitation of the in vitro polyadenylation (PA) and cleavage (CLV) reactions performed with the various substrates (Fig. 4). The data are representative of repeated experiments and have standard errors of ±6%. WT, wild type.

% WILD TYPE		
PA	CLV	
100	100	WT
107	111	LSM-A
100	118	LSM-B
97	87	LSM-1
91	75	LSM-2
73	55	LSM-3
87	57	LSM-12
60	41	LSM-13
40	30	LSM-23

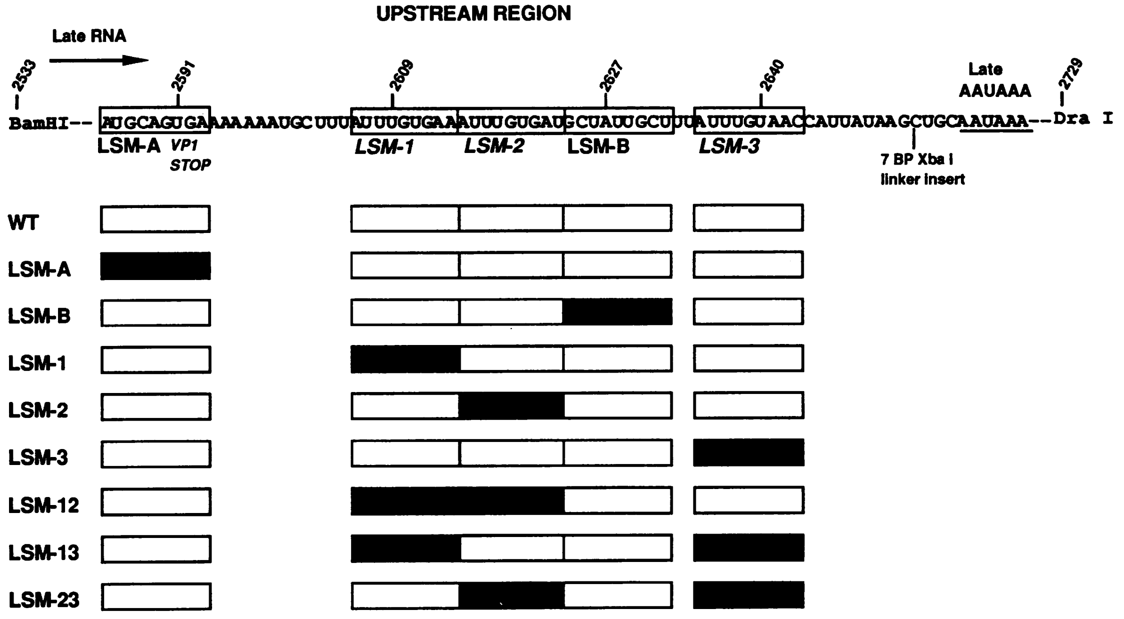


FIG. 3. Linker substitution mutants constructed with mutations in the upstream region. The positions of linker substitution mutations are indicated by boxes around the substituted sequences. In the set of mutations shown below the sequence, a white box indicates that the wild-type (WT) SV40 sequences are present while a black box indicates that the sequences have been replaced by linkers (see Materials and Methods). Mutations LSM-A and LSM-B are in sequences that do not affect the core elements indicated by deletion analysis. Mutations LSM-1, LSM-2, and LSM-3 are in the core elements 1, 2, and 3 indicated by deletion mutagenesis (Fig. 2 and 4). The table on the left shows the quantitation of the in vitro polyadenylation (PA) and cleavage (CLV) reactions performed with the linker substitution mutant substrates (Fig. 5 and 6). The data for the element 1, 2, and 3 single and double mutations were derived from a single, internally consistent experiment in which all of the substrates were simultaneously tested. However, these data are representative of repeated experiments and have standard errors of ±6%.

USEs and DSEs (7). Utilization of the polyadenylation signal containing these sites was equivalent to utilization of the wild-type signal in vivo (7).

The construction, by BAL 31 digestion, of a set of transient expression plasmids bearing deletions of various portions of the sequences upstream of the late AAUAAA hexanucleotide has been previously described (7). The deletions begin at the *Bam*HI site (SV40 nucleotide 2533), and the deleted regions are replaced by *Bam*HI linkers. Digestion with *Bam*HI and *Eco*RI produced fragments encompassing the deletions, the AAUAAA, and downstream sequences of the late poly(A) signal. These were inserted into the *Bam*HI and *Eco*RI sites of pGEM2, forming pGEM2-PA12, pGEM2-PA17, and pGEM2-PA16, which bear deletions of SV40 nucleotides 2533 to 2609 (dl2609), 2533 to 2627 (dl2627), and 2533 to 2640 (dl2640), respectively (Fig. 2).

Mutants containing a single linker substitution mutation were constructed by using a polymerase chain reaction-directed technique (35, 36). Each linker substitution mutation exactly replaced 9 bp of the wild-type sequence in pGEM-PAS with a *Bgl*II linker (CAAGATCTG). Briefly, two external primers, common to all mutants, were synthesized; one was complementary to the SP6 promoter sequences located 5′ of the *Eco*RI site, and the other was complementary to the T7 promoter sequences located 3′ of the *Bam*HI site. For each mutant, a specific set of internal primers was synthesized. The primer for each 5′-half product and the primer for each 3′-half product included CGC, 9 bases making up the *Bgl*II site, and 17 to 30 bases complementary to wild-type sequences. By using the matched external and internal primers, and pGEM2-PAS as a template, two half products were synthesized via polymerase chain reaction. The 5′-half product was cleaved with *Eco*RI and *Bgl*II; the 3′-half product was cleaved with *Bam*HI and *Bgl*II; each was gel purified. The two fragments were ligated and cut with *Bam*HI and *Eco*RI, and the products were separated on agarose gels. The appropriate dimer containing both (5′ and 3′) halves was isolated and ligated into pGEM2-PAS in place of the wild-type sequences between the *Bam*HI and *Eco*RI sites. The mutations were verified by sequencing. The resulting single linker-scanning mutants (Fig. 3) are as follows: pGEM2-UC1, which has a *Bgl*II linker substituted for SV40 nucleotides 2585 to 2593 (LSM-A); pGEM2-UM1, which has a *Bgl*II linker substituted for SV40 nucleotides 2606 to 2614 (LSM-1) and also a single base change from A to G at nucleotide 2555; pGEM2-UM2, which has a *Bgl*II linker substituted for SV40 nucleotides 2615 to 2623 (LSM-2); pGEM2-UC2, which has a *Bgl*II linker substituted for SV40 nucleotides 2624 to 2632 (LSM-B); and pGEM2-UM3, which has a *Bgl*II linker substituted for SV40 nucleotides 2635 to 2643 (LSM-3).

Linker substitution mutants pGEM2-UM12 and pGEM2-UM13 have, in addition to the *Bgl*II linker substituted for SV40 nucleotides 2606 to 2614, a *Sma*I linker (CACCCGGGT) substituted for SV40 nucleotides 2615 to 2623 (LSM-12) and 2635 to 2643 (LSM-13), respectively (Fig. 3). Double mutant pGEM2-UM32 has, in addition to the *Bgl*II linker substituted for SV40 nucleotides 2635 to 2643, a *Sma*I linker in place of SV40 nucleotides 2615 to 2623 (LSM-23) (Fig. 3). These plasmids were constructed by the polymerase chain reaction strategy outlined above, except that pGEM2-UM1 was used as the template for the construction of pGEM2-UM12 and pGEM2-UM13 and pGEM2-UM3 was used as the template for pGEM2-UM32. Likewise, the sequences of the internal primers used were based on pGEM2-UM1 and pGEM2-UM3.

The triple linker substitution mutant pGEM2-UM123 was prepared by substituting the *Bam*HI-to-*Sma*I fragment of pGEM2-UM12 for the *Bam*HI-to-*Sma*I fragment of pGEM2-UM23. This resulted in a plasmid containing *Bgl*II linkers substituting for SV40 nucleotides 2606 to 2614 and 2635 to 2643 and a *Sma*I linker substituting for nucleotides 2615 to 2623.

Preparation of polyadenylation substrate RNA. Templates for in vitro transcription of polyadenylation substrate RNAs were linearized with *Dra*I. Reactions were performed under standard conditions (2) with T7 polymerase (Promega Biotec), 50 μCi of [^{32}P]UTP (Amersham), 250 ng of linearized template, and 0.5 mM ^{7}Me-GTP (Pharmacia), which allows 5′ capping of synthesized RNAs. ^{32}P-labeled RNAs were extracted with phenol-chloroform-isoamyl alcohol (50:49:1), ethanol precipitated, and purified by electrophoresis through a 5% polyacrylamide–7 M urea gel. RNAs were eluted from gel slices in 20 mM Tris (pH 7.5)–400 mM NaCl–0.1% sodium dodecyl sulfate (SDS) at room temperature overnight. After extractions with phenol-chloroform-isoamyl alcohol and chloroform-isoamyl alcohol (49:1), RNAs were ethanol precipitated. Incorporated [^{32}P]UTP was quantitated by liquid scintillation counting.

Nuclear extracts and in vitro polyadenylation and cleavage reactions. HeLa cell nuclear extracts for in vitro polyadenylation were prepared as described by Moore (24). Polyadenylation reaction mixtures contained (final concentrations) 60% (vol/vol) nuclear extract, 1 mM ATP (Pharmacia), 20 mM phosphocreatine (Sigma), 2.6% polyvinyl alcohol, and approximately 5 fmol of ^{32}P-labeled substrate RNA in a total volume of 25 μl. Cleavage conditions were the same, except that the ATP concentration was reduced to 250 μM and 1 mM cordycepin (Boehringer Mannheim) was added. Reactions were performed at 30°C for 30 to 60 min and stopped by the addition of 20 mM Tris (pH 7.5)–400 mM NaCl–0.1% SDS to 400 μl. Reaction products were extracted once with phenol-chloroform-isoamyl alcohol, ethanol precipitated, and analyzed on 40-cm-long 5% polyacrylamide–7 M urea gels.

Substrate RNAs and polyadenylated and cleaved products were quantitated with a Molecular Dynamics PhosphorImager. Quantitated values for linker substitution and deletion mutant RNAs were adjusted to account for the U content. Percent processed RNA was calculated as the amount of processed RNA divided by the amount of processed plus unprocessed RNA.

RESULTS

In vitro polyadenylation of deletion mutant substrates. Substrate RNAs were transcribed in vitro by using T7 RNA polymerase (see Materials and Methods). Templates included the wild-type polyadenylation signal template, representing the entire SV40 late polyadenylation signal (from the *Bam*HI site at nucleotide 2533 to the *Dra*I site at nucleotide 2729) (Fig. 1), as well as deletion mutant templates dl2609, dl2627, and dl2640. These deletion mutant templates contain the wild-type sequences except for SV40 nucleotides 2533 (*Bam*HI site) to 2609, 2533 to 2627, and 2533 to 2640, respectively (Fig. 1 and 2). All substrates contain the AAUAAA and the DSEs as well as various amounts of the upstream region. Table 1 shows the expected size of each of these substrates.

TABLE 1. Sizes of in vitro polyadenylation substrates and cleavage products

Source	Size (nucleotides) of:	
	Substrate	Cleavage product
Wild type	249	195
Linker substitution mutants	249	195
dl2609	181	127
dl2627	163	109
dl2640	150	96

Figure 4 shows the results of the polyadenylation (lanes 5 to 8) and cleavage (lanes 9 to 12) reactions with these substrates. Since only 56 nucleotides are cleaved from the substrates and between 150 and 200 A residues are added, the polyadenylated species migrate above the unprocessed precursor. The expected sizes of the cleavage products formed in the presence of cordycepin are given in Table 1, and their positions of migration are shown in Fig. 4.

The polyadenylation and cleavage reactions both indicate that the efficiency of polyadenylation is progressively dimin-

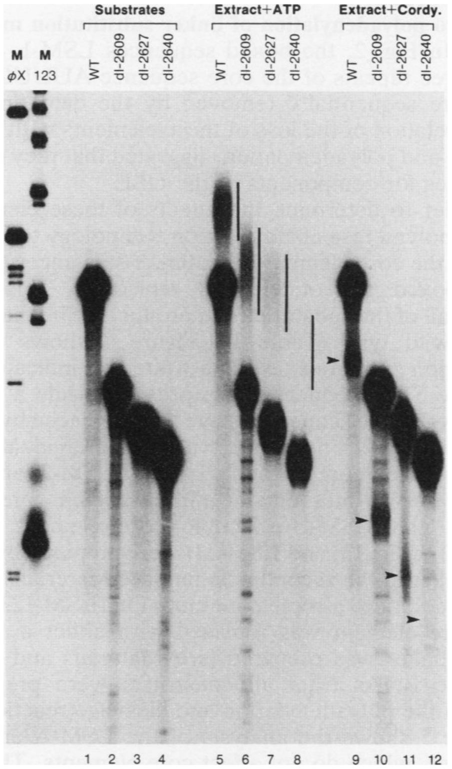


FIG. 4. In vitro polyadenylation and cleavage reactions using wild-type (WT) and deletion mutant substrate RNAs. The ^{32}P-labeled substrate RNAs (diagrammed in Fig. 2) were tested in in vitro polyadenylation (extract plus ATP) and cleavage (extract plus cordycepin [cordy.]) reactions and analyzed on 5% polyacrylamide–7 M urea gels. The sizes of the substrates and the expected cleavage products are given in Table 1. Lanes 1 to 4, migration of unprocessed substrate RNAs; lanes 5 to 8, results of in vitro polyadenylation reactions; lanes 9 to 12, results of in vitro cleavage reactions. Vertical lines, positions of migration of the polyadenylated species; arrowheads, positions of migration of the expected cleavage products. Quantitation of these and other data resulted in the values shown in Fig. 2. Lanes M, molecular size markers φX174 (φX) and 123-bp ladder.

ished as the upstream region is deleted. Figure 2 shows the results of quantitation of the percent polyadenylation and percent cleavage for each deletion mutant relative to the data for the wild type using a Molecular Dynamics Phosphorimager. Quantitative differences between the cleavage and polyadenylation reactions may arise from nonspecific polyadenylation at the ends of uncleaved substrates, which is known to occur and would skew the data upward for the polyadenylation reaction. In addition, in cases of very low levels of polyadenylation product, for example, dl2640 (Fig. 4, lane 8), we quantitated an extended region above the precursor; in our experience, the quantitation of extended regions tends to overestimate the signal, which, again, would skew the data upward. Overall, however, the data suggest that sequences within the upstream region of the SV40 late polyadenylation signal affect the efficiency of utilization of the site in vitro. These data are in substantial agreement with the results obtained by using the same mutations in transfection experiments (7). Although dl2640 has the most significant effect on cleavage efficiency, it is possible that the absolute effect noted (i.e., no detectable cleavage product) may be exaggerated because of detection problems caused by the small size of the substrate and the small size of the cleavage product.

In vitro polyadenylation of linker substitution mutant substrates. In Fig. 2, the boxed sequences LSM-1, -2, and -3 show three repeats of the core sequence AUUUGU(G/A)A which are sequentially removed by the deletion mutants. The correlation of the loss of these elements with decreased cleavage and polyadenylation suggested that they were good candidates for components of the USE.

In order to determine the effects of these elements, we utilized polymerase chain reaction technology to replace the bases of the core elements with linker sequences (Fig. 2 and 3 [the boxed nucleotides were replaced]). This approach allowed all of the substrates and products to be the same size as the wild type (Table 1). Figure 3 shows the linker substitution mutations tested; a white box indicates that the wild-type SV40 sequences are present, while a black box indicates that the sequences have been replaced by the linker sequences. *Bgl*II linkers were used to individually replace each suspected core element (LSM-1, LSM-2, and LSM-3) as well as two sites outside the suspected core elements (LSM-A and LSM-B). Double core element mutations (LSM-12, LSM-13, and LSM-23) were prepared by placing a *Sma*I linker in the second position (see Materials and Methods). In addition, a triple core mutation (LSM-123) in which each core element was replaced with either a *Sma*I or a *Bam*HI linker was prepared (see Materials and Methods). Substrate RNAs from all constructs were prepared and tested in the polyadenylation and cleavage reactions.

Figure 5 shows the analysis of the LSM-A and LSM-B substrates, which do not affect core elements. These mutations were not suspected to negatively affect polyadenylation efficiency. This is substantiated by the quantitation of the polyadenylation and cleavage data relative to those for the wild type, which is shown in Fig. 3. The polyadenylation reaction for the LSM-B mutant (Fig. 5, lane 10) appears to be less efficient than that for the wild type; this is an error due to the loading of less sample. Quantitation of these (Fig. 3) and other (not shown) data shows that LSM-A and LSM-B mutants consistently have polyadenylation and cleavage efficiencies equal to or greater than those of the wild type. Cleavage efficiency of the LSM-B substitution mutant averages at least 12% greater than that of the wild type in repeated experiments.

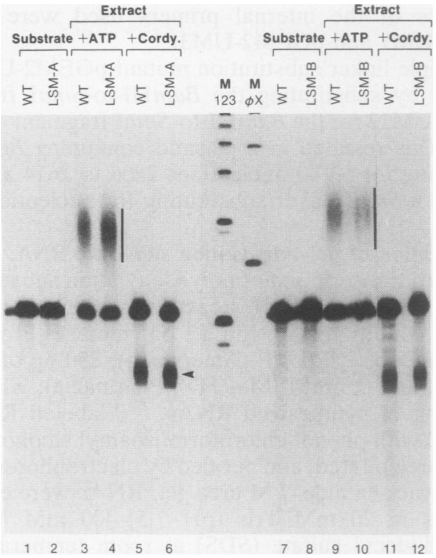


FIG. 5. In vitro cleavage and polyadenylation reactions using LSM-A and LSM-B substrate RNAs. Wild-type (WT) or mutant ^{32}P-labeled substrate RNAs were added to in vitro polyadenylation (extract plus ATP; lanes 3, 4, 9, and 10) and cleavage (extract plus cordycepin [cordy.]; lanes 5, 6, 11, and 12) reaction mixtures and analyzed on 5% polyacrylamide–7 M urea gels. Vertical lines, positions of migration of the polyadenylated RNAs; arrowheads, positions of the cleavage products. Lanes 1, 2, 7, and 8, migration of the unprocessed substrate RNAs. Quantitation of these and similar data resulted in the values shown in Fig. 3. Lanes M, molecular size markers φX174 (φX) and 123-bp ladder.

Analysis of the core element mutants is shown in Fig. 6, and quantitation of these data, relative to those for the wild type, is shown in Fig. 3. Once again, because of the nonspecific end polyadenylation of unprocessed substrate in the polyadenylation reaction, the cleavage data are considered to provide a better indication of the effect of each mutation. Linker substitution mutations of single elements indicated that the mutations had progressively greater effects as elements closer to the AAUAAA were mutated. LSM-1 had the least effect, LSM-2 had a greater effect, and LSM-3 had the greatest effect of single element mutations (cleavage efficiencies, 87, 75, and 55%, respectively, of that of the wild type). That elements closer to the AAUAAA have greater effects is in agreement with the progressive loss of efficiency noted with the deletion mutants.

Analyses of the double core element mutants confirmed and broadened the data for the single mutations. First, the distance effect was confirmed. Specifically, in LSM-12 the only wild-type core element is in position 3, in LSM-13 it is in position 2, and in LSM-23 it is in position 1. Thus, this set of mutations essentially moves a single element away from the AAUAAA, resulting in a progressive decrease in polyadenylation efficiency (57, 41, and 30%, respectively, of the wild-type level in the cleavage reaction). In addition, comparison of the single and double core element linker substitution mutants shows that the effects of multiple elements are essentially additive, i.e., the sum of the elements' individual effects relative to their distances from the AAUAAA.

Linker substitutions in core elements 1, 2, and 3 (LSM-123) showed cleavage and polyadenylation efficiencies essentially identical to those of LSM-23 (data not shown). This

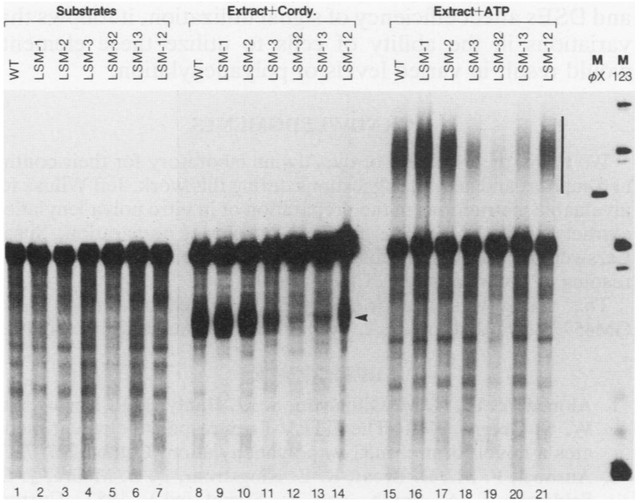


FIG. 6. In vitro cleavage and polyadenylation reactions using LSM-1, LSM-2, and LSM-3 substrate RNAs. Wild-type (WT) or mutant ^{32}P-labeled substrate RNAs were added to in vitro polyadenylation (extract plus ATP; lanes 15 to 21) and cleavage (extract plus cordycepin [cordy]; lanes 8 to 14) reaction mixtures and analyzed on 5% polyacrylamide–7 M urea gels. Vertical line, position of migration of the polyadenylated RNA; arrowhead, position of the cleavage product. Lanes 1 to 7, migration of the unprocessed substrate RNAs. Quantitation of these data resulted in the values shown in Fig. 3. Lanes M, molecular size markers ϕX174 (ϕX) and 123-bp ladder.

	CORE	CORE TO AAUAAA (Nucs.)
LSM-1	5′ UUU AUUUGU**GA AAUUUGU** 3′	44
LSM-2	**GAA AUUUGU**GA UGCUAUU	35
LSM-3	UUU AUUUGUAA CCAUUAU	15
CONSENSUS:		
-CORE ELEMENT	AUUUGURA	
-EXTENDED	uuu AUUUGURA nnnUunU	
CaMV-A	UGU AUUUGUAU UUGUAAA	19
CaMV-B	UGU AUUUGUAA AAUACUU	13
GSHV-PS1B	AUU AUUUGUAU UAGGA	
HIV	CAGCUGCUUUUUGCCUGUACUGGGUCUCUCUGGUUA	56

FIG. 7. Comparisons of the upstream region core elements and surrounding sequences. Core elements defined by LSM-1, LSM-2, and LSM-3 are aligned along with surrounding sequences on the 5′ and 3′ sides. Note that the core elements defined by LSM-1 and LSM-2 are adjacent; hence, the surrounding sequences on both sides of the core element partially overlap the adjacent element (overlaps are boldface). The core consensus and extended consensus sequences are concluded from the SV40 sequences only. Distances (in nucleotides) from the first nucleotide upstream of the AAUAAA to and including the 3′-most nucleotide of the core element are shown on the right. Sequences of the USEs defined for CaMV (19, 32), ground squirrel hepatitis virus (GSHV) (27, 28), and HIV (33) are also shown. The CaMV elements overlap (underlined regions) (see text for further discussion).

result agrees with the conclusions that core element 1 is the least effective and apparently requires one or more core elements downstream (i.e., closer to the AAUAAA) in order to provide an effect on processing.

DISCUSSION

RNA sequence elements other than AAUAAA that affect the efficiency of polyadenylation have been defined in many studies with numerous genes. Most of these studies have detected sequences lying downstream of the AAUAAA (4, 5, 10, 11, 14, 16, 17, 21–23, 29–31). It appears that some sort of downstream efficiency element, between approximately 15 and 70 nucleotides downstream of the AAUAAA, is a nearly universal characteristic of polyadenylation signals. However, the DSEs studied thus far do not define a specific consensus sequence but fall into a general class of elements which are GU or U rich. Some of these sites have been shown to be binding sites for proteins, possibly of the heterogeneous nuclear ribonucleoprotein family (34).

Elements lying upstream of the AAUAAA have been defined for a few systems (6–8, 12, 13, 19, 27, 28, 32, 33, 37). Studies in our laboratory previously defined the existence of a USE in the SV40 late polyadenylation signal by using deletion mutagenesis and transfection analysis (7). These studies defined an element between 13 and 48 nucleotides upstream of the AAUAAA (note that this distance is counted directly from the SV40 sequence; in our constructions, there are 7 additional nucleotides due to an inserted *Xba*I linker, as shown in Fig. 2 and 3 and explained above). However, detection of elements by using in vivo analyses of mutants has the possible drawback that sequence changes may alter RNA stability or, more remotely, promoter activity. Hence, we turned to in vitro polyadenylation reactions to determine the precise locations of USEs. These data are among the first to demonstrate that upstream efficiency elements affect cleavage and polyadenylation in vitro.

The in vitro analyses of the SV40 USE using deletion mutants (Fig. 2 and 4) defined essentially the same upstream region as was defined in vivo (7). Finer analysis of the upstream region was performed with linker substitution mutagenesis (Fig. 3, 5, and 6), which has the advantage of maintaining a constant size for the substrate RNA and the cleaved product, with the only differences in substrates being confined to a small substituted region. Three similar core elements were defined by linker substitution mutations LSM-1, LSM-2, and LSM-3 (Fig. 3 and 7). These core elements appear to make up the active components of the upstream region. Single and double core element linker substitution mutations indicated that elements closer to the AAUAAA had greater effects on efficiency than ones further away. For example, the element defined by LSM-3 lies 13 (20, counting the *Xba*I insert) nucleotides upstream of the AAUAAA and, when mutated, has the greatest effect on efficiency of cleavage (loss of 45% of the wild-type cleavage efficiency). However, the core elements defined by LSM-2 and LSM-1 lie 34 and 43 (41 and 50, counting the *Xba*I insert) nucleotides, respectively, upstream of the AAUAAA and have progressively smaller effects on cleavage efficiency (losses of 25 and 13% of the wild-type efficiency, respectively). In addition, analysis of single and double core element linker substitution mutants suggests that the effect of multiple elements is the sum of the elements' individual effects relative to their distances from the AAUAAA.

Figure 7 shows the sequences of the core elements defined by LSM-1, LSM-2, and LSM-3 as well as additional nucleotides on the 5′ and 3′ sides of each element. Note that the core elements defined by LSM-1 and LSM-2 are adjacent; hence, the additional sequence on either side of the core element partially overlaps the adjacent element. A very good consensus sequence for the core elements, AUUUGURA, is noted. Examination for consensus agreement within the sequences on both sides of the core elements suggests possible U richness on the 5′ side but little or no additional consensus on the 3′ side.

The USE of the CaMV polyadenylation signal has been defined in *Nicotiana tabacum* (32) and, as shown in Fig. 7, is surprisingly similar to the SV40 upstream core elements. The core elements of CaMV are overlapping, as shown in arrangement CaMV-A in Fig. 7, in which one core element is lined up with the SV40 elements and the overlapping one is underlined. In arrangement CaMV-B, the sequences have been arranged so that the underlined core element from CaMV-A is lined up with the SV40 core elements for comparison. The core elements of CaMV are nearly exact matches to the core element consensus. Examination of the CaMV sequences surrounding the core elements supports the suggestion that the U's on the 5′ side may be significant and further diminishes any relevance of sequences on the 3′ side. In addition, it is interesting that the distances between the core elements and the AAUAAA in CaMV are 13 and 19 nucleotides, very similar to the distance of the core element defined by LSM-3 (15 nucleotides), the element having the greatest effect on SV40 late polyadenylation signal efficiency. The ground squirrel hepatitis virus also has a USE necessary for polyadenylation efficiency, PS1B (27, 28), which is homologous to the SV40 and CaMV elements. However, this element is located further upstream from the polyadenylation site than the elements in SV40 and CaMV (27). This difference in position may reflect the unique polyadenylation site utilization in ground squirrel hepatitis virus, which relies on upstream information because of the presence of a variant hexanucleotide sequence (UAUAAA) that is essentially nonfunctional (27, 28). Overall, the high degree of conservation of these elements between plant and animal viruses suggests that they provide a significant, evolutionarily conserved function in RNA processing.

Given this conservation, it is interesting that the core element defined above is homologous to the Sm protein complex-binding site in human U1 RNA (20), suggesting the involvement of an evolutionarily conserved protein complex which has a known role in RNA processing. However, it has been shown that the SV40 upstream region can be functionally replaced by the HIV USE, as tested in transfection experiments (33). The HIV USE lies at least 56 nucleotides upstream of the AAUAAA within the sequences shown in Fig. 7 (33). However, the distance from the AAUAAA may be functionally smaller than indicated since the 3′ 18 nucleotides of the sequence shown affect the formation of the TAR loop in HIV transcripts. Hence, their effect on polyadenylation appears to be in the formation of this secondary structure, which functionally positions the USE (within the 5′ 18 nucleotides) closer to the AAUAAA (15a). Other than U richness, there is no apparent homology between the HIV USE and the core elements defined above. We suggest that this indicates the existence of a group of USEs that use different factors to provide similar overall functions.

How upstream and downstream efficiency elements mediate their effects on the polyadenylation reaction remains speculative. The best explanation is that they are binding sites for specific proteins which either directly affect the polyadenylation complex interactions at the AAUAAA or keep nonspecific binding proteins away from the region of the AAUAAA so that it can be easily accessed by the polyadenylation complex. By either mechanism, the efficiency elements have the potential to control gene expression through regulation of polyadenylation. Since both USEs and DSEs affect efficiency of signal utilization, it follows that variations in the ability of cells to utilize these elements would result in varied levels of polyadenylation.

ACKNOWLEDGMENTS

We thank the members of the Alwine laboratory for their continued support and helpful suggestions during this work, Jeff Wilusz for invaluable instructions in the preparation of in vitro polyadenylation extracts, Alex Valsamakis for help in extract preparation, Susan Carswell for helpful discussions, and Sherri Adams for critical reading of the manuscript. Cheers to all.

This work was supported by Public Health Service grant GM45773 awarded to J.C.A. by the National Institutes of Health.

REFERENCES

1. **Ahmed, Y. F., G. M. Gilmartin, S. M. Hanly, J. R. Nevins, and W. C. Greene.** 1991. The HTLV-I *rex* response element mediates a novel form of mRNA polyadenylation. Cell **64:**727–737.
2. **Ausubel, F. M., R. Brent, R. E. Kingstrom, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (ed.).** 1987. Current protocols in molecular biology. Greene Publishing and Wiley Interscience, New York.
3. **Bar-Shira, A., A. Panet, and A. Honigman.** 1991. An RNA secondary structure juxtaposes two remote genetic signals for human T-cell leukemia virus type I RNA 3′-end processing. J. Virol. **65:**5165–5173.
4. **Bhat, B. M., and W. S. M. Wold.** 1985. ATTAAA as well as downstream sequences are required for RNA 3′-end formation in the E3 complex transcription unit of adenovirus. Mol. Cell. Biol. **5:**3183–3193.
5. **Böhnlein, S., J. Hauber, and B. R. Cullen.** 1989. Identification of a U5-specific sequence required for efficient polyadenylation within the human immunodeficiency virus long terminal repeat. J. Virol. **63:**421–424.
6. **Brown, P. H., L. S. Tiley, and B. R. Cullen.** 1991. Efficient polyadenylation within the human immunodeficiency virus type 1 long terminal repeat requires flanking U3-specific sequences. J. Virol. **65:**3340–3343.
7. **Carswell, S., and J. C. Alwine.** 1989. Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. Mol. Cell. Biol. **9:**4248–4258.
8. **Charrington, J., and D. Ganem.** 1992. Regulation of polyadenylation in HIV: contribution of promoter proximity and upstream sequences. EMBO J. **11:**1513–1529.
9. **Chiou, H. C., C. Dabrowski, and J. C. Alwine.** 1991. Simian virus 40 late mRNA leader sequences involved in augmenting mRNA accumulation via multiple mechanisms, including increased polyadenylation efficiency. J. Virol. **65:**6677–6685.
10. **Cole, C. N., and T. P. Stacy.** 1985. Identification of sequences in the herpes simplex virus thymidine kinase gene required for efficient processing and polyadenylation. Mol. Cell. Biol. **5:**2104–2113.
11. **Conway, L., and M. Wickens.** 1985. A sequence downstream of AAUAAA is required for formation of simian virus 40 late mRNA 3′ termini in frog oocytes. Proc. Natl. Acad. Sci. USA **82:**3949–3953.
12. **DeZazzo, J. D., and M. J. Imperiale.** 1989. Sequences upstream of AAUAAA influence poly(A) site selection in a complex transcription unit. Mol. Cell. Biol. **9:**4951–4961.
13. **DeZazzo, J. D., J. E. Kilpatrick, and M. J. Imperiale.** 1991. Involvement of long terminal repeat U3 sequences overlapping the transcription control region in human immunodeficiency virus type 1 mRNA 3′ end formation. Mol. Cell. Biol. **11:**1624–1630.
14. **Gil, A., and N. J. Proudfoot.** 1984. A sequence downstream of AAUAAA is required for rabbit β-globin mRNA 3′-end formation. Nature (London) **312:**473–474.
15. **Gil, A., and N. J. Proudfoot.** 1987. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA formation. Cell **49:**399–406.

15a. **Gilmartin, G.** Personal communication.

16. **Hart, R. P., M. A. McDevitt, H. Ali, and J. R. Nevins.** 1985.

Definition of essential sequences and functional equivalence of elements downstream of the adenovirus E2A and the early simian virus 40 polyadenylation sites. Mol. Cell. Biol. **5**:2975–2983.

17. **Hart, R. P., M. A. McDevitt, and J. R. Nevins.** 1985. Poly(A) site cleavage in a HeLa nuclear extract is dependent on downstream sequences. Cell **43**:677–683.
18. **Heath, C. V., R. M. Denome, and C. N. Cole.** 1990. Spatial constraints on polyadenylation signal function. J. Biol. Chem. **265**:9098–9104.
19. **Irniger, S., H. Sanfaçon, C. M. Egli, and G. H. Braus.** 1992. Different sequence elements are required for function of the cauliflower mosaic virus polyadenylation site in *Saccharomyces cerevisiae* compared with in plants. Mol. Cell. Biol. **12**:2322–2330.
20. **Luhrmann, R., B. Kastner, and M. Bach.** 1990. Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. Biochim. Biophys. Acta **1087**:265–292.
21. **McDevitt, M. A., R. P. Hart, W. W. Wong, and J. R. Nevins.** 1986. Sequences capable of restoring poly(A) site function define two distinct downstream elements. EMBO J. **5**:2907–2913.
22. **McDevitt, M. A., M. J. Imperiale, H. Ali, and J. R. Nevins.** 1984. Requirement of a downstream sequence for generation of a poly(A) addition site. Cell **37**:993–999.
23. **McLauchlan, J., D. Gaffney, J. L. Whitton, and J. B. Clements.** 1985. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini. Nucleic Acids Res. **13**:1347–1368.
24. **Moore, C. L.** 1990. Preparation of mammalian extracts active in polyadenylation. Methods Enzymol. **181**:49–74.
25. **Moore, C. L., and P. A. Sharp.** 1984. Site-specific polyadenylation in a cell-free reaction. Cell **36**:581–591.
26. **Moore, C. L., and P. A. Sharp.** 1985. Accurate cleavage and polyadenylation of exogenous RNA substrate. Cell **41**:845–855.
27. **Russnak, R.** 1991. Regulation of polyadenylation in hepatitis B viruses: stimulation by the upstream activating signal PS1 is orientation-dependent, distance-independent, and additive. Nucleic Acids Res. **19**:6449–6456.
28. **Russnak, R., and D. Ganem.** 1990. Sequences 5′ to the polyadenylation signal mediate differential poly(A) site use in hepatitis B viruses. Genes Dev. **4**:764–776.
29. **Ryner, L. C., Y. Takagaki, and J. L. Manley.** 1989. Sequences downstream of AAUAAA signals affect pre-mRNA cleavage and polyadenylation in vitro both directly and indirectly. Mol. Cell. Biol. **9**:1759–1771.
30. **Sadofsky, M., and J. C. Alwine.** 1984. Sequences on the 3′ side of hexanucleotide AAUAAA affect efficiency of cleavage at the polyadenylation site. Mol. Cell. Biol. **4**:1460–1468.
31. **Sadofsky, M., S. Connelly, J. L. Manley, and J. C. Alwine.** 1985. Identification of a sequence element on the 3′ side of AAUAAA which is necessary for simian virus 40 late mRNA 3′-end processing. Mol. Cell. Biol. **5**:2713–2719.
32. **Sanfacon, H., P. Brodmann, and T. Hohn.** 1991. A dissection of the cauliflower mosaic virus polyadenylation signal. Genes Dev. **5**:141–149.
33. **Valsamakis, A., S. Zeichner, S. Carswell, and J. C. Alwine.** 1991. The human immunodeficiency virus type 1 polyadenylation signal: a 3′ long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylation. Proc. Natl. Acad. Sci. USA **88**:2108–2112.
34. **Wilusz, J., and T. Shenk.** 1990. A uridylate tract mediates efficient heterogeneous nuclear ribonucleoprotein C protein-RNA cross-linking and functionally substitutes for the downstream element of the polyadenylation signal. Mol. Cell. Biol. **10**:6397–6407.
35. **Zaret, K. S., J. Lin, and C. M. DePersio.** 1990. Site-directed mutagenesis reveals a liver transcription factor essential for the albumin transcriptional enhancer. Proc. Natl. Acad. Sci. USA **87**:5469–5473.
36. **Zeichner, S. L., J. Y. H. Kim, and J. C. Alwine.** 1991. Linker-scanning mutational analysis of the transcriptional activity of the human immunodeficiency virus type 1 long terminal repeat. J. Virol. **65**:2436–2444.
37. **Zhang, F., and C. N. Cole.** 1987. Identification of a complex associated with processing and polyadenylation in vitro of herpes simplex virus type 1 thymidine kinase precursor RNA. Mol. Cell. Biol. **7**:3277–3286.