JOURNAL OF VIROLOGY, Nov. 1982, p. 637–646
0022-538X/82/110637-10$02.00/0

Vol. 44, No. 2

# Complete Nucleotide Sequences of Two Adjacent Early Vaccinia Virus Genes Located Within the Inverted Terminal Repetition

SUNDARARAJAN VENKATESAN,† ALAN GERSHOWITZ, AND BERNARD MOSS*

*Laboratory of Biology of Viruses, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20205*

Received 28 May 1982/Accepted 9 August 1982

The proximal part of the 10,000-base pair (bp) inverted terminal repetition of vaccinia virus DNA encodes at least three early mRNAs. A 2,236-bp segment of the repetition was sequenced to characterize two of the genes. This task was facilitated by constructing a series of recombinants containing overlapping deletions; oligonucleotide linkers with synthetic restriction sites provided points for radioactive labeling before sequencing by the chemical degradation method of Maxam and Gilbert (Methods Enzymol. 65:499–560, 1980). The ends of the transcripts were mapped by hybridizing labeled DNA fragments to early viral RNA and resolving nuclease S1-protected fragments in sequencing gels, by sequencing cDNA clones, and from the lengths of the RNAs. The nucleotide sequences for at least 60 bp upstream of both transcriptional initiation sites are more than 80% adenine · thymine rich and contain long runs of adenines and thymines with some homology to procaryotic and eucaryotic consensus sequences. The gene transcribed in the rightward direction encodes an RNA of approximately 530 nucleotides with a single open reading frame of 420 nucleotides. Preceding the first AUG, there is a heptanucleotide that can hybridize to the 3′ end of 18S rRNA with only one mismatch. The derived amino acid sequence of the protein indicated a molecular weight of 15,500. The gene transcribed in the leftward direction encodes an RNA 1,000 to 1,100 nucleotides long with an open reading frame of 996 nucleotides and a leader sequence of only 5 to 6 nucleotides. The derived amino acid sequence of this protein indicated a molecular weight of 38,500. The 3′ ends of the two transcripts were located within 100 bp of each other. Although there are adenine · thymine-rich clusters near the putative transcriptional termination sites, specific AATAAA polyadenylic acid signal sequences are absent.

Poxviruses are large DNA viruses that replicate within the cytoplasm of infected cells. This extraordinary feat is accomplished partly by incorporating a complete viral transcriptional system within the infectious particle (19, 23). Although apparently not spliced (12, 43, 45), the mRNAs generated have typical eucaryotic features including 5′ terminal cap structures (39) and 3′ polyadenylic acid [poly(A)] tails (18, 24) enabling them to use the host translational apparatus. Studies with vaccinia virus, the prototype for this group, revealed that half of the genome is transcribed early in infection (8, 14, 20, 26). Since 100 or more early genes are scattered throughout the length of the DNA (3, 10), they must have common structural features leading to their selection by the viral multisubunit RNA polymerase (1, 25, 32) or associated factors. To identify such structural elements, we have undertaken the sequencing of several early genes.

Only recently have technical advances made it possible to characterize individual vaccinia virus transcripts. Thus far, about a dozen early mRNAs mapping within and proximal to the inverted terminal repetition have been examined (12, 43, 45). The inverted terminal repetition of vaccinia virus is about 10,000 base pairs (bp) long (13, 42) and contains a novel flip-flop loop at its distal end (2), followed by two sets of tandem 70-bp repeats (44) and then coding regions for polypeptides of approximately 7,500 daltons (7.5K), 19K, and 42K. The corresponding mRNAs are about 1,000, 600, and 1,050 nucleotides long, respectively, and are synthesized in vitro by detergent-treated virus particles (36) as well as in vivo under conditions in which

† Present address: Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20205.

viral protein or DNA synthesis is prevented (11, 43). Furthermore, all three transcripts have multiple 5′ ends as judged by the isolation of cap structures containing adenine (A) and guanine (G) as the penultimate nucleotide (36); the retention of the β-phosphate of GTP in mature mRNA is evidence against nucleolytic processing of 5′ ends (36).

Limited nucleotide sequence analysis of the 7.5K polypeptide gene revealed an extremely adenine · thymine (A · T)-rich cluster of nucleotides immediatley upstream of the cap site and a tandemly repeated hexanucleotide near the 3′ end (35). In the present communication, we provide the sequence for a 2,236-bp segment of the inverted terminal repetition that includes the entire 19K and 42K polypeptide genes as well as locations of putative transcriptional initiation and termination sites.

## MATERIALS AND METHODS

**Construction of recombinant plasmids.** Recombinant pVG3, a pBR322 derivative containing a 3.4-kilobase pair (kbp) *SalI-Eco*RI fragment from the left end of the vaccinia virus genome, was derived from a previously described recombinant, pAG1 (36). A set of in vitro deletion mutants was constructed by cleaving pVG3 at the unique *Kpn*I site (see Fig. 1). After phenol extraction and ethanol precipitation, 10 μg of linearized plasmid was digested with 1.25 U of *Bal* 31 (15). Samples were removed at 1, 2, and 3 min, at which time the reactions were stopped by heating at 65°C in the presence of 0.25% sodium dodecyl sulfate–0.012 M ethylene glycol-bis(β-aminoethyl ether)-*N,N*-tetraacetic acid (EGTA)–0.02 M EDTA. After phenol-chloroform extraction and ethanol precipitation, the extent of exonucleolytic cleavage was monitored by agarose gel electrophoresis. Frayed ends remaining after this step were filled in with reverse transcriptase (2.5 U per μg of DNA) and 0.25 mM deoxynucleoside triphosphate (dNTP) (each type) at 37°C for 30 min. *Hin*dIII oligonucleotide linkers were labeled at their 5′ ends with [γ-$^{32}$P]ATP by an exchange reaction catalyzed by polynucleotide kinase (5). Linkers, at a molar ratio of 100:1, were ligated to the plasmid DNA at 4°C for 4 h, and the ligated mix was recovered by ethanol precipitation after phenol extraction. The DNA was then digested with *Hin*dIII, phenol-chloroform extracted, and rendered free of linkers by gel filtration through a Sepharose 4B column. The plasmid DNA was then incubated at a concentration of 10 μg/ml with 7.5 U of T4 DNA ligase at 4°C for 4 h. Approximately 0.2 μg of DNA was used to transform competent *Escherichia coli* K-12 HB101 cells, and colonies were selected by growth on ampicillin plates (7). The number of transformants obtained varied from 5,000 for limited *Bal* 31 digestions to 500 for extensive digestions. Recombinant plasmids were purified by a small-scale isolation procedure (6) and screened by restriction endonuclease digestions and agarose gel electrophoresis. At least 75% of the tested recombinants had a *Hin*dIII site, and the majority of them had deletions that varied from 100 to 1,300 bp. Five recombinants with overlapping deletions (D1, D3, D4, D5, and D6) were selected for DNA sequencing.

**End labeling DNA.** Restriction fragments were treated with alkaline phosphatase and labeled at the 5′ end with [γ-$^{32}$P]ATP and T4 polynucleotide kinase (22). Avian myeloblastosis reverse transcriptase or the Klenow fragment of DNA polymerase was used to add a single complementary [α-$^{32}$P]dNTP to the recessed 3′ end of restriction fragments (47). When the 3′ end was protruding, labeling was accomplished with [α-$^{32}$P]cordycepin triphosphate and terminal deoxynucleotidyltransferase (34).

**DNA sequence analysis.** End-labeled DNA fragments were cleaved with appropriate restriction endonucleases, purified by agarose gel electrophoresis, and recovered with glass powder (37) or by electrophoresis onto DEAE paper (41). The limited chemical degradation procedure of Maxam and Gilbert (22) was used for sequencing. Reaction products were resolved on 0.4-mm-thick gels with either 4, 6, 8, or 20% polyacrylamide (30 by 160 or 40 by 80 cm) in 8 M urea (31).

**cDNA sequencing.** A library of cDNA recombinants was generously provided by B. Roberts (Harvard University School of Medicine). cDNA was prepared using early RNA from cells infected with vaccinia virus (strain WR). Oligodeoxythymidylic acid [oligo(dT)] served as a primer; hairpin 5′ ends were removed with nuclease S1, and the double-stranded cDNAs were cloned in pBR322 by deoxycytidylate and deoxyguanylate tailing (B. Roberts, personal communication). We screened the library by colony hybridization (33) with a 3-kbp *Bam*HI-*Eco*RI fragment (see Fig. 1) labeled with $^{32}$P by nick translation (29). A total of 35 colonies were selected for further restriction endonuclease analyses, and 8 were sequenced after labeling the 3′ ends with [α-$^{32}$P]cordycepin triphosphate (34).

**Mapping 5′ and 3′ ends by nuclease S1 protection.** poly(A)-containing RNA was purified from the cytoplasm of cells 4 h after infection with vaccinia virus (15 PFU per cell) in the presence of cycloheximide or was made in vitro by vaccinia virus particles (36). The RNA was hybridized to DNA fragments containing a label at the 5′ or 3′ end on the coding strand (35). After nuclease S1 digestion (35), hybrids were analyzed by electrophoresis on sequencing polyacrylamide gels.

**Materials.** Restriction endonucleases were purchased from New England Biolabs (Beverly, Mass.) or Bethesda Research Laboratories (Gaithersburg, Md.). Nucleases *Bal* 31 and S1 were supplied by New England Biolabs and Miles Laboratories, Inc., Elkhart, Ind., respectively. Both terminal deoxynucleotidyltransferase and T4 polynucleotide kinase came from P-L Biochemicals, Milwaukee, Wis., whereas DNA polymerase holoenzyme and the Klenow fragment came from Boehringer Mannheim Corp. Indianapolis, Ind. T4 DNA ligase was purchased from Bethesda Research Laboratories, and avian myeloblastosis virus reverse transcriptase was supplied by J. W. Beard of Life Sciences, Coral Gables, Fla. Oligodeoxynucleotides were obtained from Collaborative Research, Inc., Waltham, Mass., [α-$^{32}$P]dNTP and [γ-$^{32}$P]ATP from Amersham Searle, Chicago, Ill., and [α-$^{32}$P]cordycepin triphosphate from New England Nuclear Corp., Boston, Mass.

## RESULTS

**Sequencing strategy.** Three mRNAs encoding polypeptides of approximately 7.5K, 19K, and

42K were mapped within the inverted terminal repetition of the vaccinia virus genome (11, 42, 43) (Fig. 1). A strategy, avoiding either extensive restriction endonuclease mapping or shotgun methods, was used to determine the genomic sequence of the 19K and 42K polypeptides. This deletion linker approach required the cloning in pBR322 of the 3.4-kbp *SalI-EcoRI* fragment located 5.6 to 9.0 kbp from the end of the viral DNA (Fig. 1). The recombinant is referred to as pVG3. A unique *KpnI* site, located within the viral DNA segment, was cleaved, and the ends of the linearized plasmid were subjected to exonuclease digestion with *Bal* 31 (15). After the addition of synthetic oligonucleotide linkers containing a *HindIII* site, the plasmids were recircularized and used to transform *E. coli*. Recombinant plasmids were screened, and a set with overlapping deletions was selected (Fig. 1). Sequencing of each recombinant was facilitated by use of the mobile *HindIII* site for 5′ and 3′ labeling. Additional sequencing was carried out using the restriction sites and conventional strategy indicated at the bottom of Fig. 1.

A sequence of 2,236 nucleotides encompassing genes for the 19K and 42K polypeptides is shown in Fig. 2. As will be discussed later, both genes have long open translational reading frames of opposite polarity.

**Location of the 5′ ends of transcripts.** The mRNAs for the 19K and 42K polypeptides are transcribed from opposite strands of the inverted terminal repetition (Fig. 1). Approximate map positions of these mRNAs were determined by use of restriction fragments for hybridization selection and cell-free translation and for probing blots of electrophoretically separated RNAs. To more precisely locate their 5′ ends, a modifi-

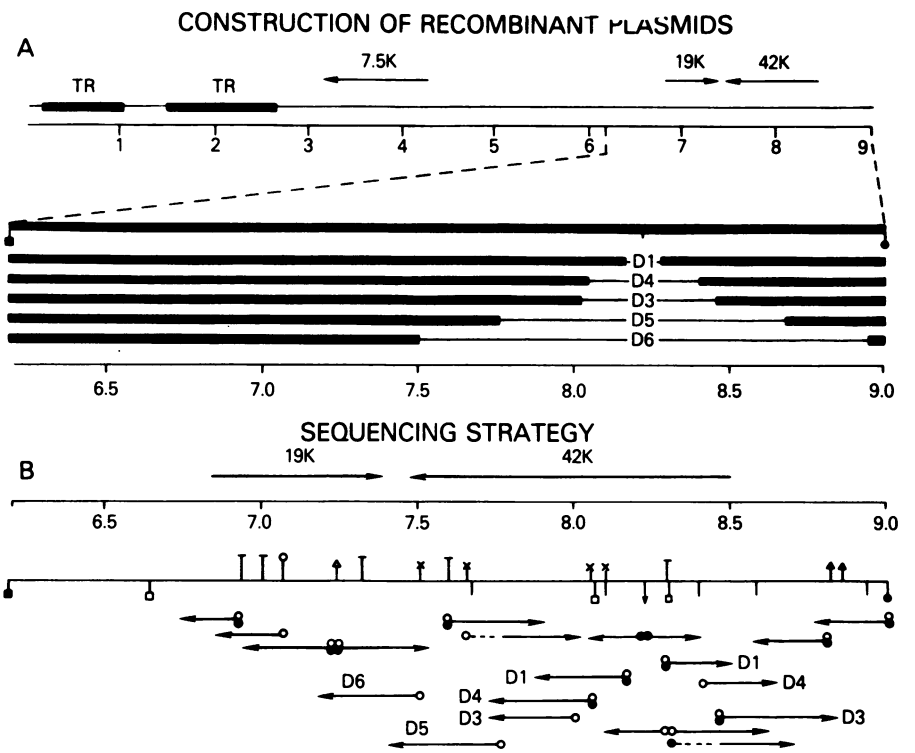


FIG. 1. Construction of recombinants and sequencing strategy. (A) Part of the 10-kbp inverted terminal repetition of vaccinia virus is shown. The dark blocks labeled TR represent the two sets of 70-bp tandem repeats. Arrows represent the approximate lengths, map positions, and directions of transcription of three early mRNAs encoding polypeptides of approximately 7.5K, 19K, and 42K. The *BamHI-EcoRI* segment of the recombinant plasmid pVG3 is expanded. Additional recombinants were constructed from pVG3 using *Bal* 31 to make deletions at the *KpnI* site. The extent of the deletions in recombinants D1, D4, D3, D5, and D6 is shown by gaps in the heavy bar. (B) The *BamHI-EcoRI* segment of the inverted terminal repetition with the mRNAs for the 19K and 42K polypeptides is shown. Symbols for relevant restriction endonuclease sites are: *AvaII*, ♀; *BamHI*, ■; *EcoRI*, ●; *HincII*, ⏐; *HinfI*, ⏐; *HpaII*, △; *KpnI*, ▼; *TaqI*, ⊤; *XbaI*, ⊻. Filled and unfilled circles represent sites of 3′ and 5′ labeling, respectively. Solid arrows indicate the extent of sequence determination; interrupted parts of the arrow indicate regions where the sequence was not determined. The designations D1 to D6 refer to the use of deletion recombinants; otherwise, pVG3 or a related recombinant pVG1 was used.

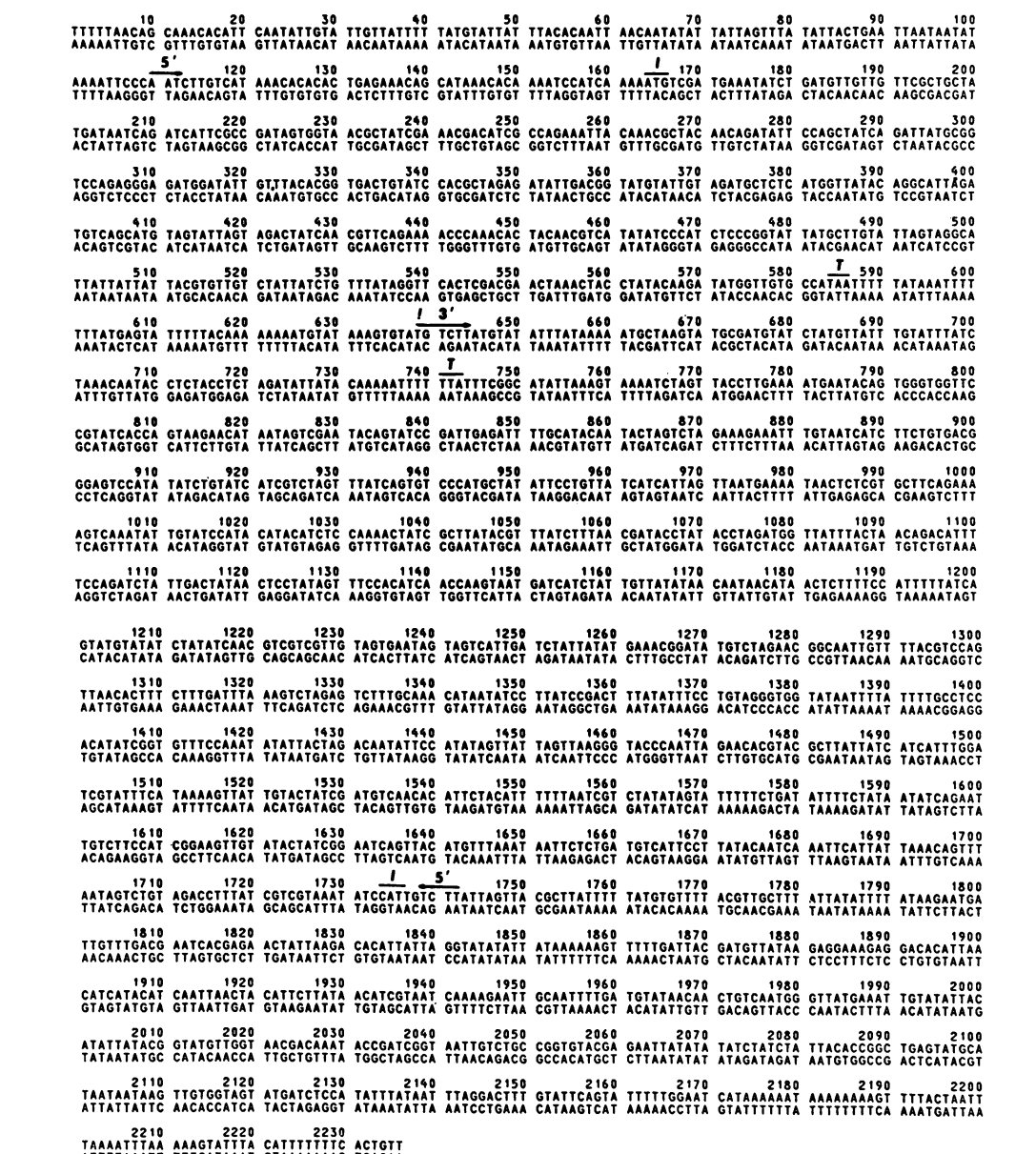

FIG. 2. Nucleotide sequence of a 2,236-bp segment of DNA encompassing the genes for the 19K and 42K polypeptides. The nucleotide numbered 1 is approximately 6,800 nucleotides from the end of the genome, and nucleotide 2,236 is a few nucleotides before the first *Eco*RI site. The upper and lower lines represent the noncoding strands for mRNAs expressing the 19K and 42K polypeptides, respectively. The putative transcriptional initiation sites for the two mRNAs are indicated by an arrow with 5′ over it. The 3′ end(s) of the mRNA for the 19K polypeptide is shown by an arrow with 3′ written over it. The 3′ end of the mRNA for the 19K polypeptide identified by cDNA sequencing is shown by!. I and T refer to the presumptive translational initiation and termination codons, respectively, of the two mRNAs.

cation introduced by Weaver and Weissman (38) of the nuclease S1 procedure of Berk and Sharp (4) was used. A *Hin*cII-*Ava*II fragment (6.6 to 7.1 kbp from the end of the genome) with 5′ $^{32}$P label at the *Ava*II site was employed. The labeled strand of this fragment was expected to contain sequences complementary to the 5′ end of the 19K polypeptide mRNA. After hybridization to early RNA, made in infected cells or in vitro by detergent-treated virus particles, re-

maining single-stranded DNA was digested with nuclease S1. The radioactively labeled DNA segment that was protected from nuclease digestion represented the distance from the RNA 5′ terminus to the 5′ end of the DNA probe. The size of this segment was determined by polyacrylamide gel electrophoresis under denaturing conditions. As shown in Fig. 3, at least six closely spaced nuclease-resistant bands were resolved. By coelectrophoresis with the DNA sequence reaction products of the original *Hinc*II-*Ava*II fragment, the nuclease-resistant bands were lined up with a sequence ladder. As pointed out previously (16), a 1.5-bp downward displacement of the nuclease S1 bands is necessary for proper alignment. This placed the 5′ ends of the RNA complementary to the sequence AGATT in Fig. 3. Since the nuclease-protected bands were spaced one nucleotide apart, the heterogeneity could result from a

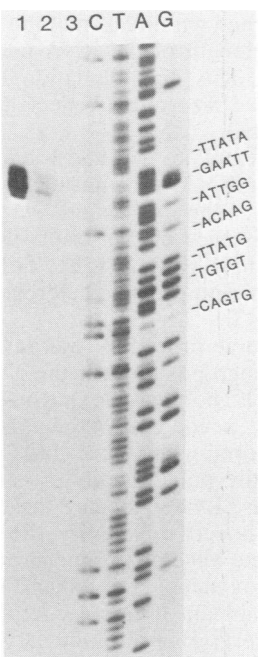


FIG. 3. Map locations of the 5′ end(s) of mRNA encoding the 19K polypeptide. Approximately 0.5 pmol of the *Ava*II-*Hinc*II fragment (7.0 to 6.64 kbp from the end of the DNA) 5′ labeled at the *Ava*II site was denatured and incubated under conditions favoring RNA-DNA hybridization with 5 to 10 pmol of poly(A) selected early in vivo RNA or in vitro RNA. After hybridization, single-stranded DNA was digested with nuclease S1 (750 U/ml). Resistant products after incubation with (1) in vivo RNA, (2) in vitro RNA, or (3) no RNA and sequence reaction products (C [cytosine]; T is T plus C; A is A plus G; G) were resolved on an 8% polyacrylamide gel (30 by 160 cm) in 8 M urea.

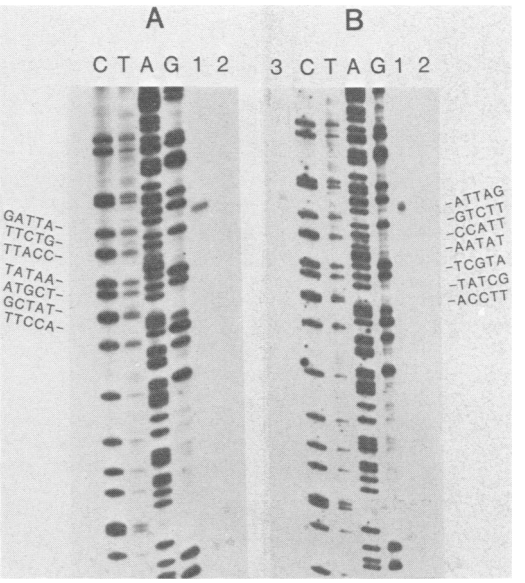


FIG. 4. Map locations of the 5′ ends of mRNA encoding the 42K polypeptide. Approximately 0.5 pmol of (A) an *Xba*I-*Eco*RI fragment (8.1 to 9.0 kbp from the end of the DNA) and (B) a *Taq*I-*Eco*RI fragment (8.3 to 9.0 kbp from the end of the DNA) 5′ labeled at the *Xba*I and *Taq*I sites, respectively, was hybridized to approximately 25 pmol of poly(A) selected early cytoplasmic RNA. Nuclease S1-resistant products obtained with (1) in vivo RNA, (2) no RNA, or (3) in vitro RNA and sequence reaction products (C; T is T plus C; A is A plus G; G) were resolved on an 8% polyacrylamide gel (30 by 160 cm) in 8 M urea.

tendency of nuclease S1 to leave overhanging ends or to nibble into a base-paired region. Analysis of the capped 5′ ends of this message indicated that 80% ended in A and 20% in G, suggesting that the major ends are complementary to one or both of the adjacent deoxythymidylate residues (positions 110 and 111 in Fig. 2.). The map position of the 5′ end was confirmed by using as a probe the *Hinc*II-*Hpa*II fragment (6.6 to 7.2 kbp from the end of the DNA) labeled at the 5′ end at the *Hpa*II site (not shown).

The 5′ ends of the message encoding the 42K polypeptide were mapped in an analogous manner. Two DNA fragments, *Xba*I-*Eco*RI (8.1 to 9.0 kbp from the end of the DNA) and *Taq*I-*Eco*RI (8.3 to 9.0 kbp from the end of the DNA), labeled at the *Xba*I and *Taq*I site, respectively, were employed. The major nuclease-resistant bands lined up within a sequence TCTT (Fig. 4). Previous cap analysis indicated approximately equal amounts of G and A ends suggesting that they are complementary to the CT residues of the above sequence (numbered 1,740 and 1,741 in Fig. 2).

**Location of the 3′ ends of transcripts.** Previous studies indicated that the mRNAs for the 19K

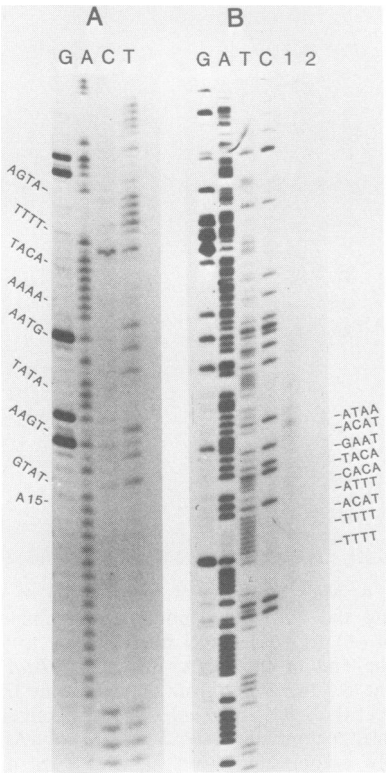


FIG. 5. Map locations of the 3′ ends of mRNA encoding the 19K polypeptide by cDNA sequencing and nuclease S1 protection. (A) A recombinant plasmid (pVBR12) was selected from a vaccinia cDNA library by its ability to hybridize specifically to a vaccinia DNA fragment encoding mRNA for the 19K polypeptide. The recombinant DNA was cleaved at the *Pst*I site and 3′ labeled with [α-$^{32}$P]cordycepin triphosphate and terminal deoxynucleotidyltransferase. After cleavage with *Hpa*II, the smaller of the labeled DNA fragments was purified, sequenced by the chemical degradation method, and analyzed by electrophoresis on 8% polyacrylamide gel. The sequence of the noncoding (i.e., RNA) strand is indicated. The 15 adenylate residues represent the proximal portion of the poly(A) tail of the mRNA. (B) Approximately 0.5 pmol of an *Hpa*II-*Hin*fI fragment (7.2 to 7.6 kbp from the end of the DNA) labeled at the 3′ end was hybridized to 10 pmol of poly(A) selected early cytoplasmic RNA in 80% formamide at 38°C. Nuclease S1-resistant products obtained with (1) or without (2) RNA and sequence reaction products (G; A is A plus G; T is T plus C; C) were resolved on an 8% polyacrylamide gel (30 by 160 cm) in 8 M urea. The coding strand DNA sequence is indicated.

and 42K polypeptides are approximately 580 and 1,050 nucleotides long, respectively (36, 43). Using the map locations determined above for the 5′ ends, this would place the 3′ ends of the oppositely oriented 19K and 42K polypeptide messages close to each other near positions 690 and 740, respectively, of the genomic sequence in Fig. 2.

Attempts were made to map the 3′ ends of both mRNAs more precisely. The first approach involved a procedure analogous to that used for mapping 5′ termini. A *Hpa*II-*Hin*fI fragment (7.2 to 7.6 kbp from the end of the DNA) labeled at the 3′ end at the *Hpa*II site was used as a probe for the 19K polypeptide message. After hybridization to early RNA, the nuclease-protected DNA segment represented the distance from the 3′ end of the RNA to the labeled restriction site. The size of this segment was determined by polyacrylamide gel electrophoresis, again using a sequence ladder prepared from the original hybridization probe for comparison. Two major and several minor bands were detected (Fig. 5B). After making the appropriate correction, they appeared to line up with the complementary AA residues of the sequence GAAT. This corresponds to the two thymine (T) residues located at positions 643 and 644 of Fig. 2 and places the 3′ end within 50 bp of the site predicted from the length of the RNA using northern blot analysis. This slight deviation from its originally reported size might reflect the length of the poly(A) tail.

Similar attempts to map the 3′ end of the 42K polypeptide mRNA were unsatisfactory, possibly because of the low abundance of the message, the existence of transcripts with overlapping 3′ ends that might compete for hybridization, and possible heterogeneity of the 3′ ends (36, 43).

An alternative procedure was investigated to confirm the map positions of the 3′ ends of the 19K and 42K polypeptide mRNAs. For this approach we needed recombinants containing cDNAs prepared using an oligo(dT) primer hybridized to the poly(A) tails of early vaccinia virus mRNAs. By sequencing the recombinant cDNAs we hoped to identify the nucleotides adjacent to the oligo(dT) primer or complementary oligodeoxyadenylate [oligo(dA)]. Such a cDNA recombinant library, generously provided by B. Roberts (Harvard Medical School), was screened by colony hybridization. Selected recombinants were further characterized by restriction endonuclease analyses, and eight were sequenced from their *Pst*I sites. The sequence of one is shown in Fig. 5A. After a cluster of deoxycytidylate residues, derived from the deoxyguanylate and deoxycytidylate tailing used for construction of the recombinants, there were 15 deoxyadenylate (dA) residues and then a sequence corresponding to the genome near the 3′ end of the 19K polypeptide. The deoxythymidylate residue adjacent to the oligo(dA) is at position 639 of Fig. 2, four nucleotides downstream from the 3′ end deduced by nuclease S1

protection experiments. Further comparison of the cDNA sequence with the genomic sequence (position 678 to 638, Fig. 2) revealed two minor discrepancies. There were seven dAs in cDNA in contrast to eight dAs in genomic DNA (position 618, Fig. 2) and six dAs in cDNA corresponding to seven dAs in the genome (position 603, Fig. 2). These differences might represent slippage during reverse transcription or minor variations in the vaccinia virus WR isolates used for genome sequencing and cDNA cloning. Since these differences occur in the nontranslated region of the gene, there might be no biological consequences of such variations.

Unfortunately, the other seven cDNA recombinants from this region of the genome as well as most from a second region lacked the oligo(dA) sequence and therefore could not be used to identify the precise 3′ end of the message.

## DISCUSSION

The majority of DNA viruses engage cellular RNA polymerases for gene transcription within the nucleus of the infected cell. Thus it is not surprising that their regulatory signals resemble those of the host (9). By contrast, poxviruses are cytoplasmic and possess a unique transcriptional system making it likely that distinctive DNA sequences interact with the viral RNA polymerase. To elucidate structural features related to the organization and expression of the vaccinia virus genome, several regions have been sequenced. These include the terminal loops that link the two DNA strands together and adjacent DNA (2) and portions of the most distal gene coding for a 7.5K polypeptide (35). To these is now added a 2,236-bp segment that encodes two early polypeptides. Sequencing was achieved by the chemical degradation method of Maxam and Gilbert (22) using a strategy that involved the preparation of a series of recombinants with overlapping deletions and inserted restriction site linkers. This deletion linker approach eliminated the need for extensive restriction site mapping or shotgun sequencing.

A procedure of Weaver and Weissman (38) was used to map the 5′ ends of both mRNAs. After hybridization of a DNA fragment labeled at the 5′ end to RNA, the nuclease S1-protected segment and the nucleotide sequence reaction products of the original fragment were compared by electrophoresis on the same polyacrylamide gel. These data and our finding of $m^7G(5')pppA^m$ and $m^7G(5')pppG^m$ caps on both the 19K and 42K polypeptide mRNAs (36) placed two major 5′ ends at or near nucleotides 110 and 111 and 1,740 and 1,741 of Fig. 2, respectively. A previous analysis of the 5′ ends of the 7.5K polypeptide mRNA also suggested the presence of multiple closely spaced purine ends (35).

Since the cap structures of both the 19K and 42K mRNAs retain the β-phosphate of GTP (36), the 5′ ends of their messages correspond to sites of transcriptional initiation. Accordingly, the adjacent DNA may contain promoter recognition sequences. In Fig. 6, regions upstream of the 5′ ends of the mRNAs coding for polypeptides estimated to be 7.5K, 19K, and 42K are shown. For each, the 60 bp preceding the start site are at least 80% A · T rich with many runs of A's and T's including some 14 to 18 nucleotides long. Embedded within the A · T-rich region are possible equivalents of the Pribnow and Hogness-Goldberg boxes of procaryotes and eucaryotes, respectively (9, 30). Nevertheless, the similarity to the eucaryotic TATA sequence is not exact. Interestingly, the sequence CGTAAAA is found starting at −28 and −29 of the 7.5K and 42K polypeptide genes, respectively (Fig. 6). In addition, each of the genes contains other homologous A · T-rich clusters including CAAT 40 to 60 bp upstream of the cap sites.

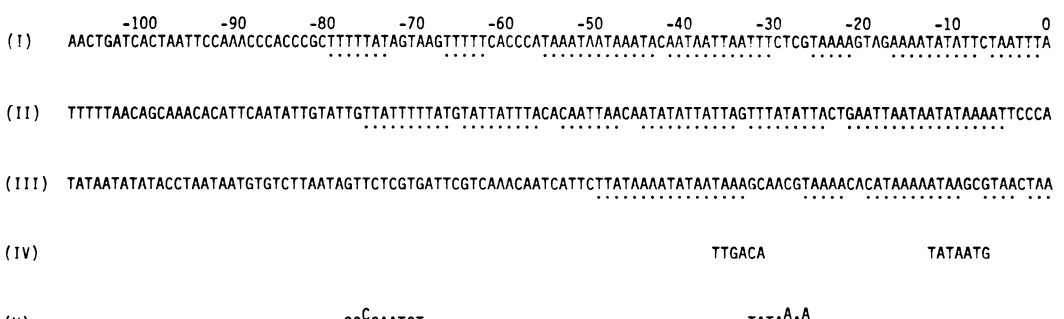


FIG. 6. Sequences upstream of three early transcriptional sites of vaccinia virus. I, II, and III refer respectively to the upstream sequences of mRNAs encoding 7.5K, 19K, and 42K polypeptides. The most proximal initiation site is indicated by O. Lines IV and V represent the near upstream and far upstream sequences in procaryotes and eucaryotes, respectively.

Examination of more genes will be necessary to know whether to designate any of the sequences referred to above as specific "pox-boxes." To assess the functional significance of DNA structures, genetic experiments are required. With this in mind, we have recently mapped the selectable thymidine kinase gene of vaccinia virus (40). Our plans are to mutate the putative promoter of the thymidine kinase gene within plasmid or phage recombinants and then to introduce the mutated DNA into infectious virus by transfection procedures. In this manner, structure/function relationships can be explored.

A sequence of four or more nucleotides preceding the translational initiation codon in procaryotic mRNA is complementary to the 3′ terminal segment of 16S rRNA. This complementarity is thought to facilitate the binding of mRNA to ribosomes. The situation in eucaryotes is less certain since the presence of a sequence complementary to 18S rRNA is variable (17, 46). For this reason, it was of interest to examine the leader sequences of early vaccinia virus mRNAs. Before the first ATG of the 19K polypeptide gene, there is a heptanucleotide ATCCATC that has close sequence complementarity to the 3′ end of 18S rRNA (3′AUUAC*UAGGAAG*GGCG, indicated in italics; 17). Similarly, a complementary tetranucleotide TCCT was noted (35) a few nucleotides before the second ATG of the 7.5K polypeptide gene. However, less homology to the 3′ end of 18S rRNA is present in the 42K polypeptide gene which appears to have a leader of only five to six nucleotides. Whether these differences in leader sequences are responsible for the greater in vitro synthesis of 19K and 7.5K polypeptides

```
                                        193                                 223                                 253
ATG TCG ATG AAA TAT CTG ATG TTG TTG TTC GCT GCT ATG ATA ATC AGA TCA TTC GCC GAT AGT GGT AAC GCT ATC GAA ACG ACA TCG CCA
MET SER MET LYS TYR LEU MET LEU LEU PHE ALA ALA MET ILE ILE ARG SER PHE ALA ASP SER GLY ASN ALA ILE GLU THR THR SER PRO

                                        283                                 313                                 343
GAA ATT ACA AAC GCT ACA ACA GAT ATT CCA GCT ATC AGA TTA TGC GGT CCA GAG GGA GAT GGA TAT TGT TTA CAC GGT GAC TGT ATC CAC
GLU ILE THR ASN ALA THR THR ASP ILE PRO ALA ILE ARG LEU CYS GLY PRO GLU GLY ASP GLY TYR CYS LEU HIS GLY ASP CYS ILE HIS

                                        373                                 403                                 433
GCT AGA GAT ATT GAC GGT ATG TAT TGT AGA TGC TCT CAT GGT TAT ACA GGC ATT AGA TGT CAG CAT GTA GTA TTA GTA GAC TAT CAA CGT
ALA ARG ASP ILE ASP GLY MET TYR CYS ARG CYS SER HIS GLY TYR THR GLY ILE ARG CYS GLN HIS VAL VAL LEU VAL ASP TYR GLN ARG

                                        463                                 493                                 523
TCA GAA AAC CCA AAC ACT ACA ACG TCA TAT ATC CCA TCT CCC GGT ATT ATG CTT GTA TTA GTA GGC ATT ATT ATT ATT ACG TGT TGT CTA
SER GLU ASN PRO ASN THR THR THR SER TYR ILE PRO SER PRO GLY ILE MET LEU VAL LEU VAL GLY ILE ILE ILE ILE THR CYS CYS LEU

                                        553                                 583
TTA TCT GTT TAT AGG TTC ACT CGA CGA ACT AAA CTA CCT ATA CAA GAT ATG GTT GTG CCA TAA
LEU SER VAL TYR ARG PHE THR ARG ARG THR LYS LEU PRO ILE GLN ASP MET VAL VAL PRO END

MOLECULAR WEIGHT = 15525

                                       1707                                1677                                1647
ATG GAT ATT TAC GAC GAT AAA GGT CTA CAG ACT ATT AAA CTG TTT AAT AAT GAA TTT GAT TGT ATA AGG AAT GAC ATC AGA GAA TTA TTT
MET ASP ILE TYR ASP ASP LYS GLY LEU GLN THR ILE LYS LEU PHE ASN ASN GLU PHE ASP CYS ILE ARG ASN ASP ILE ARG GLU LEU PHE

                                       1617                                1587                                1557
AAA CAT GTA ACT GAT TCC GAT AGT ATA CAA CTT CCG ATG GAA GAC AAT TCT GAT ATT ATA GAA AAT ATC AGA AAA ATA CTA TAT AGA CGA
LYS HIS VAL THR ASP SER ASP SER ILE GLN LEU PRO MET GLU ASP ASN SER ASP ILE ILE GLU ASN ILE ARG LYS ILE LEU TYR ARG ARG

                                       1527                                1497                                1467
TTA AAA AAT GTA GAA TGT GTT GAC ATC GAT AGT ACA ATA ACT TTT ATG AAA TAC GAT CCA AAT GAT GAT AAT AAG CGT ACG TGT TCT AAT
LEU LYS ASN VAL GLU CYS VAL ASP ILE ASP SER THR ILE THR PHE MET LYS TYR ASP PRO ASN ASP ASP ASN LYS ARG THR CYS SER ASN

                                       1437                                1407                                1377
TGG GTA CCC TTA ACT AAT AAC TAT ATG GAA TAT TGT CTA GTA ATA TAT TTG GAA ACA CCG ATA TGT GGA GGC AAA ATA AAA TTA TAC CAC
TRP VAL PRO LEU THR ASN ASN TYR MET GLU TYR CYS LEU VAL ILE TYR LEU GLU THR PRO ILE CYS GLY GLY LYS ILE LYS LEU TYR HIS

                                       1347                                1317                                1287
CCT ACA GGA AAT ATA AAG TCG GAT AAG GAT ATT ATG TTT GCA AAG ACT CTA GAC TTT AAA TCA AAG AAA GTG TTA ACT GGA CGT AAA ACA
PRO THR GLY ASN ILE LYS SER ASP LYS ASP ILE MET PHE ALA LYS THR LEU ASP PHE LYS SER LYS LYS VAL LEU THR GLY ARG LYS THR

                                       1257                                1227                                1197
ATT GCC GTT CTA GAC ATA TCC GTT TCA TAT AAT AGA TCA ATG ACT ACT ATT CAC TAC AAC GAC GAC GTT GAT ATA GAT ATA CAT ACT GAT
ILE ALA VAL LEU ASP ILE SER VAL SER TYR ASN ARG SER MET THR THR ILE HIS TYR ASN ASP ASP VAL ASP ILE ASP ILE HIS THR ASP

                                       1167                                1137                                1107
AAA AAT GGA AAA GAG TTA TGT TAT TGT TAT ATA ACA ATA GAT GAT CAT TAC TTG GTT GAT GTG GAA ACT ATA GGA GTT ATA GTC AAT AGA
LYS ASN GLY LYS GLU LEU CYS TYR CYS TYR ILE THR ILE ASP ASP HIS TYR LEU VAL ASP VAL GLU THR ILE GLY VAL ILE VAL ASN ARG

                                       1077                                1047                                1017
TCT GGA AAA TGT CTG TTA GTA AAT AAC CAT CTA GGT ATA GGT ATC GTT AAA GAT AAA CGT ATA AGC GAT AGT TTT GGA GAT GTA TGT ATG
SER GLY LYS CYS LEU LEU VAL ASN ASN HIS LEU GLY ILE GLY ILE VAL LYS ASP LYS ARG ILE SER ASP SER PHE GLY ASP VAL CYS MET

                                        987                                 957                                 927
GAT ACA ATA TTT GAC TTT TCT GAA GCA CGA GAG TTA TTT TCA TTA ACT AAT GAT GAT AAC AGG AAT ATA GCA TGG GAC ACT GAT AAA CTA
ASP THR ILE PHE ASP PHE SER GLU ALA ARG GLU LEU PHE SER LEU THR ASN ASP ASP ASN ARG ASN ILE ALA TRP ASP THR ASP LYS LEU

                                        897                                 867                                 837
GAC GAT GAT ACA GAT ATA TGG ACT CCC GTC ACA GAA GAT GAT TAC AAA TTT CTT TCT AGA CTA GTA TTG TAT GCA AAA TCT CAA TCG GAT
ASP ASP ASP THR ASP ILE TRP THR PRO VAL THR GLU ASP ASP TYR LYS PHE LEU SER ARG LEU VAL LEU TYR ALA LYS SER GLN SER ASP

                                        807                                 777                                 747
ACT GTA TTC GAC TAT TAT GTT CTT ACT GGT GAT ACG GAA CCA CCC ACT GTA TTC ATT TTC AAG GTA ACT AGA TTT TAC TTT AAT ATG CCG
THR VAL PHE ASP TYR TYR VAL LEU THR GLY ASP THR GLU PRO PRO THR VAL PHE ILE PHE LYS VAL THR ARG PHE TYR PHE ASN MET PRO

AAA TAA
LYS END

MOLECULAR WEIGHT = 38508
```

FIG. 7. Derived amino acid sequences of the two early polypeptides. Open reading frames are shown with their predicted molecular weights.

relative to the 42K polypeptide (11) is not understood.

Nuclease S1 protection experiments were also used to map the major 3′ ends of the 19K polypeptide mRNA at or near nucleotides 643 and 644 (Fig. 2). We considered that a more precise identification of the sequence adjacent to the poly(A) tail of the message could be obtained by sequencing cDNA recombinants. Indeed, a T residue at position 639 was found adjacent to a nongenomic oligo(dA) cluster. The sequence of additional cDNA recombinants is necessary to determine the extent of 3′ heterogeneity. However, as mentioned above, most of the cDNA recombinants did not have the oligo(dA) · oligo(dT) terminus because the second polymerase reaction did not go to completion or this structure was nibbled away during the nuclease S1 step used to remove the 5′ hairpin. An alternative method of cDNA cloning that avoids this step would seem preferable (21). For a variety of possible reasons, we could not precisely locate the 3′ end of the 42K polypeptide message. However, from the location of the 5′ end and the length of the RNA, it must occur between nucleotides 640 and 740. Thus, the 3′ ends of the two oppositely oriented mRNAs must be very close to each other.

Examination of the genomic sequence near the 3′ ends of the two mRNAs does not reveal the tandem CTATTC that was found with the 7.5K polypeptide message (35). Evidently, this repeated structure is not a general termination sequence. Just before the end of the 19K polypeptide mRNA there is a possibly related sequence CTTATG. The significance of the tandemly repeated sequence TAGAGGTAGAGG beyond the coding regions of the 42K polypeptide mRNA is questionable (Fig. 2). Although A · T-rich sequences are present, a precise AATAAA poly(A) signal sequence (27) was not found. It is not known whether the 3′ ends of vaccinia virus mRNAs represent termination sites or sites of RNA processing.

The genomic sequence in Fig. 2 was analyzed with the aid of a computer program (28) for amino acid and stop codons in both directions and all three reading frames. The first ATG occurs at position 164, about 50 nucleotides downstream from the start site for the 19K polypeptide mRNA. A second ATG occurs inphase, six nucleotides further downstream. Between the first ATG and the TAA stop codon at position 584, there is a 420-nucleotide open reading frame sufficient to code for 140 amino acids (Fig. 2). Additional inphase stop codons occur before the RNA terminates. In both other reading frames, there are many stop codons throughout the gene.

On the opposite DNA strand, an ATG occurs at position 1,736, only five to six nucleotides downstream from the 5′ end of the 42K polypeptide mRNA. A 993-nucleotide open reading frame continues until the TAA at position 743 near the end of the message. This region would code for a polypeptide of 331 amino acids (Fig. 2). Both of the other reading frames have stop codons throughout.

Based on the predicted amino acid sequence (Fig. 7), the molecular weights of the two polypeptides would be 15.5K and 38.5K. These numbers are somewhat lower than the 19K and 42K values determined by polyacrylamide gel electrophoresis (11). It seems likely that the latter values were overestimated since in that study the endogenous reticulocyte 43K polypeptide appeared to be 48K.

The amino acid sequences of the two polypeptides are unremarkable. Neither has a cluster of hydrophobic amino acids indicative of a membrane transport function. A · T-rich codons are used primarily, reflecting the 60% A · T content of the genes. As yet, there is no information regarding the functions of these two early proteins.

## LITERATURE CITED


1. **Baroudy, B. M., and B. Moss.** 1980. Purification and characterization of a DNA-dependent RNA polymerase from vaccinia virions. J. Biol. Chem. **255**:4372–4380.
2. **Baroudy, B. M., S. Venkatesan, and B. Moss.** 1982. Incompletely base-paired flip-flop terminal loops link the two DNA strands of the vaccinia virus genome into one uninterrupted polynucleotide chain. Cell **28**:315–324.
3. **Belle Isle, H., S. Venkatesan, and B. Moss.** 1981. Cell-free translation of early and late mRNAs selected by hybridization to cloned DNA fragments derived from the left 14 million to 72 million daltons of the vaccinia virus genome. Virology **112**:306–317.
4. **Berk, A. J., and P. A. Sharp.** 1977. Sizing and mapping of early adenovirus mRNA's by gel electrophoresis of S1 endonuclease-digested hybrids. Cell **12**:721–732.
5. **Berkner, K. L., and W. R. Folk.** 1977. Polynucleotide kinase exchange reaction. Quantitative assay for restriction endonuclease-generated 5-phosphoryl termini in DNAs. J. Biol. Chem. **252**:3176–3184.
6. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1513–1523.
7. **Bolivar, F., R. L. Rodriguez, P. J. Greene, M. E. Betlach, H. L. Heyneker, and H. W. Boyer.** 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2**:95–112.
8. **Boone, R. F., and B. Moss.** 1978. Sequence complexity and relative abundance of vaccinia virus mRNA's synthesized in vivo and in vitro. J. Virol. **26**:554–569.
9. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
10. **Cabrera, C. V., M. Esteban, R. McCarron, W. T. McAllister, and J. A. Holowczak.** 1978. Vaccinia virus transcription: hybridization of mRNA to restriction fragments of vaccinia DNA. Virology **86**:102–114.
11. **Cooper, J. A., R. Wittek, and B. Moss.** 1981. Hybridization selection and cell-free translation of mRNA's encoded within the inverted terminal repetition of the vaccinia virus genome. J. Virol. **37**:284–294.
12. **Cooper, J. A., R. Wittek, and B. Moss.** 1981. Extension of

the transcriptional and translational map of the left end of the vaccinia virus genome to 21 kilobase pairs. J. Virol. **39**:733–745.

13. **Garon, C. F., E. Barbosa, and B. Moss.** 1978. Visualization of an inverted terminal repetition in vaccinia virus DNA. Proc. Natl. Acad. Sci. U.S.A. **75**:4863–4867.
14. **Grady, L. J., and E. Paoletti.** 1977. Molecular complexity of vaccinia DNA and the presence of reiterated sequences in the genome. Virology **79**:337–341.
15. **Gray, H. B., Jr., D. A. Ostrander, J. L. Hodnett, R. J. Legerski, and D. L. Robberson.** 1975. Extracellular nucleases of pseudomonas Bal 31. Characterization of single strand-specific deoxyriboendonuclease and doublestrand deoxyriboexonuclease activities. Nucleic Acids Res. **2**:1459–1492.
16. **Green, M. R., and R. G. Roeder.** 1980. Definition of a novel promoter for the major adenovirus-associated virus mRNA. Cell **22**:231–242.
17. **Hagenbüchle, O., M. Santer, J. A. Steitz, and R. J. Mons.** 1978. Conservation of the primary structure at the 3′ end of 18S mRNA from eukaryotic cells. Cell **13**:551–563.
18. **Kates, J. R., and J. Beeson.** 1970. Ribonucleic acid synthesis in vaccinia virus. II. Synthesis of polyriboadenylic acid. J. Mol. Biol. **50**:19–33.
19. **Kates, J. R., and B. R. McAuslan.** 1967. Poxvirus DNA-dependent RNA polymerase. Proc. Natl. Acad. Sci. U.S.A. **58**:134–141.
20. **Kaverin, N. V., N. L. Varich, V. V. Surgay, and V. I. Chernos.** 1975. A quantitative estimation of poxvirus genome fraction transcribed as "early" and "late" mRNA. Virology **65**:112–119.
21. **Land, H., M. Grez, H. Hauser, W. Lindenmaier, and G. Schutz.** 1981. 5′-terminal sequences of eukaryotic mRNA can be cloned with high efficiency. Nucleic Acids Res. **9**:2251–2266.
22. **Maxam, A., and W. Gilbert.** 1980. Sequencing end labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
23. **Munyon, W., E. Paoletti, and J. T. Grace, Jr.** 1967. RNA polymerase activity in purified infectious vaccinia virus. Proc. Natl. Acad. Sci. U.S.A. **58**:2280–2287.
24. **Nevins, J. R., and W. K. Joklik.** 1975. Poly(A) sequences of vaccinia virus messenger RNA: nature, mode of addition and function during translation *in vitro* and *in vivo*. Virology **63**:1–14.
25. **Nevins, J. R., and W. K. Joklik.** 1977. Isolation and properties of the vaccinia virus DNA-dependent RNA polymerase. J. Biol. Chem. **252**:6930–6938.
26. **Oda, K., and W. K. Joklik.** 1967. Hybridization and sedimentation studies on "early" and "late" vaccinia messenger RNA. J. Mol. Biol. **27**:395–419.
27. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ noncoding region sequences in eukaryotic mRNA. Nature (London) **263**:211–214.
28. **Queen, C. L., and L. J. Korn.** 1980. Computer analysis of nucleic acids and proteins. Methods Enzymol. **65**:595–609.
29. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113**:237–251.
30. **Rosenberg, M., and D. Court.** 1980. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
31. **Sanger, F., and R. Coulson.** 1978. The use of thin acrylamide gels for DNA sequencing. FEBS Lett. **87**:107–110.
32. **Spencer, E., S. Shuman, and J. Hurwitz.** 1980. Purification and properties of vaccinia virus DNA-dependent RNA polymerase. J. Biol. Chem. **255**:5388–5395.
33. **Thayer, R. E.** 1979. An improved method for detecting foreign DNA in plasmids of *Escherichia coli*. Anal. Biochem. **98**:60–63.
34. **Tu, C.-P., and S. N. Cohen.** 1980. 3′-end labeling of DNA with [$\alpha$-$^{32}$P] cordycepin triphosphate. Gene **10**:177–183.
35. **Venkatesan, S., B. M. Baroudy, and B. Moss.** 1981. Distinctive nucleotide sequences adjacent to multiple initiation and termination sites of an early vaccinia virus gene. Cell **25**:805–813.
36. **Venkatesan, S., and B. Moss.** 1981. In vitro transcription of the inverted terminal repetition of the vaccinia virus genome: correspondence of initiation and cap sites. J. Virol. **37**:738–747.
37. **Vogelstein, B., and D. Gillespie.** 1979. Preparative and analytical purification of DNA from agarose. Proc. Natl. Acad. Sci. U.S.A. **76**:615–619.
38. **Weaver, R. F., and C. Weissman.** 1979. Mapping of RNA by a modification of the Berk-Sharp procedure: the 5′ termini of $\beta$-globin mRNA have identical map coordinates. Nucleic Acids Res. **7**:1175–1193.
39. **Wei, C.-M., and B. Moss.** 1975. Methylated nucleotides block 5′-terminus of vaccinia virus messenger RNA. Proc. Natl. Acad. Sci. U.S.A. **72**:318–322.
40. **Weir, J. P., G. Bajszár, and B. Moss.** 1982. Mapping of the vaccinia virus thymidine kinase gene by marker rescue and by cell-free translation of selected mRNA. Proc. Natl. Acad. Sci. U.S.A. **79**:1210–1214.
41. **Winberg, G., and M. L. Hammarskjold.** 1980. Isolation of DNA from agarose gels using DEAE-paper. Application to restriction site mapping of adenovirus type 16. Nucleic Acids Res. **8**:253–264.
42. **Wittek, R., E. Barbosa, J. A. Cooper, C. F. Garon, H. Chan, and B. Moss.** 1980. The inverted terminal repetition in vaccinia virus DNA encodes early mRNAs. Nature (London) **285**:21–25.
43. **Wittek, R., J. A. Cooper, E. Barbosa, and B. Moss.** 1980. Expression of the vaccinia virus genome: analysis and mapping of mRNAs encoded within the inverted terminal repetition. Cell **21**:487–493.
44. **Wittek, R., and B. Moss.** 1980. Tandem repeats within the inverted terminal repetition of vaccinia virus DNA. Cell **21**:277–284.
45. **Wittek, R., and B. Moss.** 1982. Colinearity of RNAs with the vaccinia virus genome; anomalies with two complementary early and late RNAs result from a small deletion or rearrangement within the inverted terminal repetition. J. Virol. **42**:447–455.
46. **Yamaguchi, K., S. Hikaka, and K.-I. Miura.** 1982. Relationship between structure of the 5′ noncoding region of viral mRNA and efficiency in the initiation step of protein synthesis in a eukaryotic system. Proc. Natl. Acad. Sci. U.S.A. **79**:1012–1016.
47. **Yang, R. C. A., and R. Wu.** 1979. BK virus DNA sequence coding for the amino-terminus of the T-antigen. Virology **92**:340–352.