Proc. Natl. Acad. Sci. USA
Vol. 76, No. 2, pp. 565–569, February 1979
Biochemistry

Nucleotide sequence of the BK virus DNA segment encoding small t antigen

(papovaviruses/small t and large T antigens/termination codon/splice sites/amino acid sequence)

RAVI DHAR, ISABELLE SEIF, AND GEORGE KHOURY

Laboratory of Molecular Virology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014

Communicated by Paul C. Zamecnik, September 29, 1978

ABSTRACT **The nucleotide sequence from 0.64 to 0.53 map units in the BK virus genome coding for the small t protein has been determined. There is only one open reading frame that can code for a polypeptide of 172 amino acids, the putative small t protein. Beyond this segment, multiple termination codons are present in all three reading frames. There is considerable nucleotide and amino acid sequence homology between this region of BK virus and the analogous region of simian virus 40, especially in the proximal portion from 0.64 to 0.60 map units which is most likely common to the small t and large T BK virus proteins. A comparison of the conserved sequences within the early papovavirus genes both confirms the evolutionary relationship between these viruses and suggests the amino acid composition of the regions required for T antigen functions.**

BK virus (BKV) is a human papovavirus originally isolated from the urine of renal transplant recipients (1) and subsequently from the urines of a number of immunosuppressed patients (2). The isolation of BKV from humans, the presence of anti-BKV antibodies in 70% of most adult populations (3), and the preferential growth of this virus in human cells in tissue culture constitute evidence that BKV is a human virus, distinct from simian virus 40 (SV40). Like polyomavirus and SV40, the BKV protein capsid contains a double-stranded superhelical genome with a molecular weight of about 3.4×10^6. A number of studies have demonstrated a similarity between the genomes of SV40 and BKV in both genetic organization and nucleotide sequence homology (4–6). Although the early regions of these two viruses appear dissimilar, hybridization techniques with lowered stringency have provided evidence for approximately 25% mismatch in this portion of BKV and SV40 genomes (ref. 7; P. M. Howley, unpublished results).

Although the late BKV capsid polypeptides are easily distinguished from those of SV40 on the basis of size and antigenicity, BKV produces an early T antigen that strongly cross-reacts immunologically with SV40 (1, 8–10). Recent experiments indicate that the BKV T antigen can substitute functionally in a lytic infection for a defective SV40 T antigen (ref. 11; C.-J. Lai, N. Goldman, and G. Khoury, unpublished observations).

A number of studies now indicate that there are at least two forms of SV40 and BKV T antigens, small-t and large-T, which have approximate molecular weights of 17,000 and 94,000, respectively (12–15). These polypeptides are encoded by two distinct early viral mRNAs which differ in their size and splicing pattern (16–18). As a consequence, large-T and small-t share amino-terminal sequences and are distinct at their carboxy-terminal ends. Large T antigen has been shown to play a crucial role in the initiation of viral DNA replication, the regulation of transcription, and in cell transformation (see ref. 19). Recent studies suggest that the small t antigen may also play a role in transformation (20, 21). The complex manner in which the mRNAs for these early papovavirus peptides are processed and regulated indicates that a number of important control regions, including splice sites, are situated within the proximal half of the early region. Because of the functional similarity between the BKV and SV40 large T antigens and the location of control regions within the SV40 and BKV early regions, it was of interest to comparatively analyze their nucleotide sequences. This study presents the nucleotide sequence of the proximal portion of the early BKV region which appears to encode the BKV small t antigen.

MATERIALS AND METHODS

[γ-^{32}P]ATP (specific activity 2000–3500 Ci/mmol) (1 Ci = 3.7×10^{10} becquerels) was purchased from ICN Chemical and Radioisotope. The restriction enzymes were obtained from New England BioLabs or Bethesda Research Laboratories (Rockville, MD), T4 polynucleotide kinase was purchased from P-L Biochemicals, and snake venom phosphodiesterase was obtained from Boehringer Mannheim.

Viral DNA. Secondary cultures of human embryonic kidney cells were infected with prototype BKV (1, 22) at a multiplicity of 1–5 plaque forming units/cell. After 3 days, viral DNA was purified from the Hirt supernatant fraction by described methods (22).

Direct DNA Sequencing. Prototype BKV DNA is cleaved by *Hin*dIII into four fragments, which can be purified by electrophoresis on 4% polyacrylamide gels (23). The sequence of the *Hin*dIII-C fragment from 0.625 to 0.645 map units has been reported elsewhere (22). The *Hin*d-III-D and B fragments either were directly labeled at the 5′ end with ^{32}P or were recleaved with *Ava* II, *Mbo* II, *Alu* III, or *Hae* III endonucleases and then labeled at their 5′ ends. Appropriate restriction endonucleases were used to cleave these 5′-end-labeled fragments, generating DNA segments labeled at only one end. The end-labeled fragments were purified by polyacrylamide gel electrophoresis and direct DNA sequencing was performed as described by Maxam and Gilbert (24). A few nucleotides near the 5′ end were determined by partial digestion with snake venom phosphodiesterase and subsequent two-dimensional homochromatography (25).

RESULTS AND DISCUSSION

BKV DNA Sequence. In this study, we have extended the previously determined nucleotide sequence for the origin of BKV DNA replication (22) into the early gene region including the segment that appears to encode the BKV small t antigen. The sequence of nucleotides 1–108 (Fig. 2) which maps within the *Hin*d-III-C fragment near the *Hin*dIII-C/D junction

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: BKV, BK virus; SV40, simian virus 40.

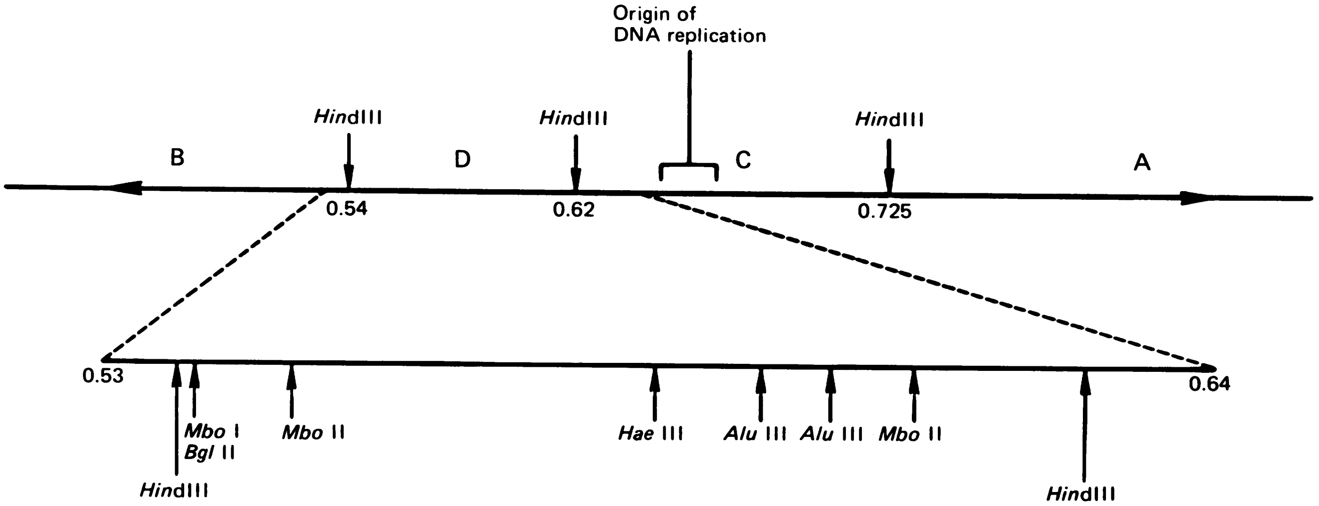

FIG. 1. Cleavage map of the BKV genome based on *Hin*dIII restriction endonuclease sites. The region from 0.53 to 0.64 map units is enlarged and vertical arrows indicate the cleavage sites for additional restriction endonucleases.

(Fig. 1) has been published elsewhere (22). The analysis of the remainder of the sequence between map positions 0.62 and 0.53 was expedited by the presence of a large number of restriction enzyme cleavage sites in this region as shown in Fig. 1. Most of the sequence presented in this study was confirmed by direct DNA sequencing of the complementary DNA strand. The largest single uncleaved stretch was a 180-nucleotide fragment from nucleotide 304 to 483 (Fig. 2). The sequence of part of this fragment was obtained from only one DNA strand, especially near the ends of the restriction endonuclease cleavage sites; the sequence of the rest of the genomic segment presented in this study was obtained from both strands by using the Maxam and Gilbert procedure (24). Near the junctions of the restriction endonuclease cleavage sites, nucleotide sequences were confirmed by analysis of the overlapping fragments. The first few nucleotides near the junctions were also confirmed by partial digestion of 5′-end ^{32}P-labeled fragments with snake venom phosphodiesterase and two-dimensional homochromatography (25).

Location and Deduced Sequence of the Early BKV Proteins. As in SV40, two early viral proteins have been identified in human cells infected by BKV (15). The large T and small t antigens have apparent molecular weights of 90,000 and 17,000, respectively, similar to the sizes of the analogous SV40 proteins (12–15). If the mRNA for SV40 large T antigen were a continuous transcript of the early region, the polypeptide would be interrupted by termination codons that are present in all three reading frames (17, 18, 26). Recent results from transcriptional mapping studies have shown that the two major early mRNAs for small t and large T antigens are spliced (16–18). The transcript for the SV40 small t antigen is spliced on the 3′ side of the termination codon and thus can encode a polypeptide of 174 amino acids from the initiator AUG at map position 0.647 (nucleotides 5081–5079) to the first termination codon at map position 0.547 (nucleotides 4559–4557; Fig. 2; refs. 17 and 18).

The SV40 large T mRNA is spliced between map positions 0.534–0.600 (14, 16, 17), which removes the termination codons at 0.547 map unit as well as the coding sequences for the 92 carboxy-terminal amino acids that appear to be present in small t antigen. This splice has the consequence of uniting the proximal portion of the large T mRNA with an open reading frame that extends from 0.534 map unit to the carboxyl terminus of large T antigen at 0.174 map unit (17, 18). Recently Paucha *et al.* (14) have determined a partial amino-terminal amino acid sequence of the SV40 T antigens. These are in complete agreement with the amino acids predicted from the SV40 DNA sequence, confirming the assignment of initiator AUG and reading frame. The first AUG triplet on the early BKV coding strand is located at nucleotides 10–12 (Fig. 2) and appears to be analogous to the SV40 initiation signal. This initiator codon introduces an open reading frame for the putative BKV small t antigen, spanning nucleotides 10 to 526 or 172 potential codons, followed by the termination codon, UAA.

The proposed amino acid sequence for BKV t antigen is presented below the nucleotide sequence in Fig. 2. The similarity with the SV40 small t antigen, presented in the same figure, is striking. Because the BKV polypeptide is presumably two amino acids shorter, two gaps (XXX) have been inserted into the nucleotide sequence at positions 236–238 and 256–258 to optimize the homology between the two viral DNAs. Based on the degree of similarity, the coding sequence for the proposed BKV t antigen seems to be naturally divided into two regions—the first from nucleotides 10 to 234 and the second from nucleotides 235 to 525. Extensive homology between BKV and SV40 occurs in the first region encoding the amino-terminal portion of both large T and small t antigens. In this region, 178 of 225 nucleotides (79%) and 65 of 75 amino acids (87%) are identical. In 30 cases an altered triplet (26 of these in the third base position) codes for the same amino acid in BKV and SV40; in many cases in which the amino acid is changed in this region, the substituted amino acid is similar in its properties (e.g., Leu for Ile, Arg for Lys, and Asp for Glu). As will be mentioned below, the proximal large T splice site is probably located near the junction of these two segments of the small t protein (near nucleotide position 250–254; Fig. 2). Thus the region of highly conserved structure appears to be that segment common to small t and large T antigens. With this kind of sequence homology, it is perhaps not surprising that the BKV large T antigen can substitute for a defective SV40 large T antigen. For example, BKV preinfection of monkey cells restores a number of biological properties in the early SV40 tsA mutant grown at nonpermissive temperature, including the initiation of DNA replication, the regulation of viral transcription, and the "helper function" for growth of adenovirus in monkey cells (ref. 11; C. J. Lai, N. Goldman, and G. Khoury, unpublished results).

The second region of BKV small t antigen (nucleotides 235–525) diverges somewhat more from SV40 (nucleotides 4865–4560). In this second region, 192 out of 291 nucleotides (66%), but only 56 of 97 amino acids (58%) are identical. An

```
                 10           20           30            40           50           60            70           80           90
                  |            |            |             |            |            |             |            |            |
BKV   TTT GCA AAA ATG GAT AAA GTT CTT AAC AGG GAA GAA TCC ATG GAG CTC ATG GAC CTT TTA GGC CTT GAA AGA GCT GCC TGG GGA AAT CTT
                  met asp lys val leu asn arg glu glu ser met glu leu met asp leu leu gly leu glu arg ala ala trp gly asn leu
                                                          leu gln                                     ser                 ile
SV40  TTT GCA AAG ATG GAT AAA GTT TTA AAC AGA GAG GAA TCT TTG CAG CTA ATG GAC CTT CTA GGT CTT GAA AGG AGT GCC TGG GGG AAT ATT
      |                                       |                                       |
     5090                                    5060                                    5030

                 100          110          120          130          140          150          160          170          180
                  |            |            |             |            |            |             |            |            |
      CCC TTA ATG AGA AAA GCT TAT TTA AGG AAG TGT AAG GAA TTT CAC CCT GAC AAA GGG GGC GAC GAG GAC AAA ATG AAG AGA ATG AAT ACT
      pro leu met arg lys ala tyr leu arg lys cys lys glu phe his pro asp lys gly gly asp glu asp lys met lys arg met asn thr
                                      lys                                                         glu             lys
      CCT CTG ATG AGA AAG GCA TAT TTA AAA AAA TGC AAG GAG TTT CAT CCT GAT AAA GGA GGA GAT GAA GAA AAA ATG AAG AAA ATG AAT ACT
      |                                       |                                       |
     5000                                    4970                                    4940

                 190          200          210          220          230              240           250              260
                  |            |            |             |            |                |             |                |
      TTG TAT AAA AAA ATG GAG CAG GAT GTA AAG GTA GCT CAT CAG CCT GAT TTT GGA XXX ACC TGG AGT AGC TCA GAG GTT XXX TGT GCT GAT
      leu tyr lys lys met glu gln asp val lys val ala his gln pro asp phe gly xxx thr trp ser ser ser glu val xxx cys ala asp
                              asp gly         tyr                             gly phe     asp ala thr         phe ala ser ser
      CTG TAC AAG AAA ATG GAA GAT GGA GTA AAA TAT GCT CAT CAA CCT GAC TTT GGA GGC TTC TGG GAT GCA ACT GAG GTA TTT GCT TCT TCC
      |                                       |                                       |
     4910                                    4880                                    4850

           270        280          290          300           310          320          330           340          350
            |             |            |            |             |            |            |             |            |
      TTT CCT CTT TGC CCA GAT ACC CTG TAC TGC AAG GAA TGG CCT GAG TGT TCC AAA AAG CCT TCT GTG CAC TGC CCT TGC ATG CTA TGT CAG
      phe pro leu cys pro asp thr leu tyr cys lys glu trp pro glu cys ser lys lys pro ser val his cys pro cys met leu cys gln
      leu asn pro gly val     ala met             gln                 ala         met     ala asn     ile     leu         leu
      TTA AAT CCT GGT GTT GAT GCA ATG TAC TGC AAA CAA TGG CCT GAG TGT GCA AAG AAA ATG TCT GCT AAC TGC ATA TGC TTG CTG TGC TTA
      |                                       |                                       |
     4820                                    4790                                    4760

           360        370          380          390           400          410          420           430          440
            |             |            |            |             |            |            |             |            |
      CTT AGA TTA AGG CAT TCA AAT AGA AAA TTT TTA AGA AAA GAG CCC TTG GTT TGG ATA GAT TGC TAC TGC ATT GAC TGC TTC ACA CAG TGG
      leu arg leu arg his ser asn arg lys phe leu arg lys glu pro leu val trp ile asp cys tyr cys ile asp cys phe thr gln trp
              met lys     glu             leu tyr         asp                 val                 phe             arg met
      CTG AGG ATG AAG CAT GAA AAT AGA AAA TTA TAC AGG AAA GAT CCA CTT GTG TGG GTT GAT TGC TAC TGC TTC GAT TGC TTT AGA ATG TGG
      |                                       |                                       |
     4730                                    4700                                    4670

           450        460          470          480           490          500          510           520          530
            |             |            |            |             |            |            |             |            |
      TTT GGC TTA GAC CTA ACT GAA GAA ACT CTG CAA TGG TGG GTC CAA ATA ATT GGA GAA ACT CCC TTC AGA GAT CTA AAG CTT TAA GGT AAC
      phe gly leu asp leu thr glu glu thr leu gln trp trp val gln ile ile gly glu thr pro phe arg asp leu lys leu
                          cys     gly         leu leu     cys asp                 gln     thr tyr
      TTT GGA CTT GAT CTT TGT GAA GGA ACC TTA CTT CTG TGG TGT GAC ATA ATT GGA CAA ACT ACC TAC AGA GAT TTA AAG CTA TAA GGT AAA
      |                                       |                                       |
     4640                                    4610                                    4580

           540        550          560
            |             |            |
      TAA CTT ATA TTT AGA TAA ATA ATA AAA

      TAT AAA ATT TTT AAG TGT ATA ATG TGT
      |
     4550
```

FIG. 2. Nucleotide sequence of a portion of the early BKV genome that codes for the small t protein. The BKV nucleotide sequence from 0.64 to 0.53 map units is shown on the upper line. On the lower line is an analogous sequence of the SV40 genome from 0.65 to 0.54 map units (17, 18). Between the two sets of sequences are the putative amino acids for the BKV and SV40 small t polypeptides. Because the BKV protein is two amino acids shorter, two sites marked XXX are inserted at positions that least disturb the comparative homology.

Position 1	Position 2				Position 3
	U	C	A	G	
U	Phe 5	Ser 1	Tyr 2	Cys 4	U
	2	2	2	7	C
	Leu 6	2	Ochre 1	Opal	A
	2	0	Amber	Trp 7	G
C	Leu 7	Pro 6	His 2	Arg 0	U
	1	3	2	0	C
	3	1	Gln 2	0	A
	2	0	4	0	G
A	Ile 2	Thr 4	Asn 3	Ser 1	U
	0	2	1	1	C
	2	1	Lys 9	Arg 7	A
	Met 8	0	7	3	G
G	Val 3	Ala 4	Asp 7	Gly 0	U
	1	1	6	3	C
	2	0	Glu 7	3	A
	1	0	7	1	G

FIG. 3. Codon utilization for the BKV small t polypeptide.

outstanding feature of this region is its abundance of cysteine residues (8 of 10 in the same position in BKV and SV40). Our data suggest that, although both of the BKV small t regions share extensive nucleotide homology with SV40, the amino acid sequence is more highly conserved in the first region (nucleotides 10–234).

Nine methionine-containing tryptic peptides have been observed for the BKV small t protein; eight of these are common to the BKV large T polypeptide (15). The nucleotide sequence of the BKV small t region presented in Fig. 2 predicts seven methionine-containing tryptic peptides, six of which would be shared with the large-T of BKV if splicing occurs, as suggested, around nucleotide 250—i.e., at a position analogous to the splice site in the large T mRNA of SV40 (16–18). Four methionine-containing tryptic peptides, common to both small and large T antigens, were shown to be identical in BKV and SV40 (15). Our data suggest three methionine-containing tryptic peptides are identical, including the amino-terminal peptide Met-Asp-Lys. Thus, although there are small discrepancies between the sequence predictions and empirical results, these sets of data are basically in agreement.

Regulatory Regions. In addition to the putative initiator codon AUG for the BKV large T and small t antigens, and the terminator codon UAA for the small t protein, the sequence presented in this manuscript presumably contains at least two splice sites for the early BKV mRNAs. Recently, analogous splice sites in SV40 have been elucidated by using the method of primer extension and reverse transcription across the spliced RNA junction (27). By comparison with the proximal (donor) splice site for the SV40 large T mRNA (approximately 4837) it would appear that the BKV site (nucleotides 250–254) is identical (. . . GAGGT), suggesting that this sequence may be specifically recognized. Also, by analogy with SV40, the proximal or donor site for small t antigen mRNA is presumably located just after the BKV small t termination codon (. . . TAA GGT . . .; nucleotides 526–531). The acceptor splice for the 3′ segment of both BKV early RNAs is most likely present just beyond the 3′ end of the sequence presented in this manuscript (unpublished data). Although the recognition signals for putative splicing enzymes are not known, the trinucleotide AGG has been recognized in the vicinity of many splice sites (17, 27). The similarities in sequence between the regions coding for the BKV and SV40 early transcripts suggest that similar enzymes recognize splice sites in both viral RNAs. Just beyond the proximal small t splice site are several termination codons in all three reading frames (Fig. 2); these are presumably removed by the splicing events to generate the mRNA for the BKV large T antigen.

Codon Utilization. The presumed distribution and composition of codons for BKV small t antigen are similar to the nonrandom use of codons for amino acids described for the SV40 genome (17, 18). The deficiency of the dinucleotide CG in rabbit and human β-globin mRNA (28, 29) and in SV40 (17, 18) is also found in BKV (Fig. 3). Thus, none of the Arg codons of the form CGX are used for BKV small-t. Additional codons that are absent from the BKV t gene include CCG, ACG, AUC, GCG, and UCG, none of which are found in any of the SV40 coding sequences (17, 18). The nonrandom nature of code words is presumably dependent on such cell-related factors as t-RNA distribution and mRNA structures. It is therefore not surprising that the primate papovavirus codons would have a similar composition.

We thank C.-J. Lai, J. Alwine, S. Segal, and N. Goldman for helpful discussions during the course of these studies; C.-J. Lai, P. Howley, D. Simmons, and H. Hoeksema for their critical review of the manuscript; and M. Lewis and M. König for excellent technical assistance.

1. Gardner, S. D., Field, A. M., Coleman, D. V. & Hulme, B. (1971) *Lancet* **i**, 1253–1257.
2. Takemoto, K. K., Rabson, A. S., Mullarkey, M. F., Blaese, R. M., Garon, C. F. & Nelson, D. (1974) *J. Natl. Cancer Inst.* **53**, 1205–1207.
3. Shah, K. V., Daniel, R. N. & Worszawski, S. (1973) *J. Infect. Dis.* **128**, 784–787.
4. Howley, P. M., Mullarkey, M. F., Takemoto, K. K. & Martin, M. A. (1975) *J. Virol.* **15**, 173–181.
5. Gardner, S. D. (1973) *Br. Med. J.* **i**, 77–78.
6. Osborn, J. E., Robertson, S. M., Padgett, B. L., Walker, D. L. & Weisblum, B. (1976) *J. Virol.* **19**, 675–684.
7. Newell, N., Lai, C.-J., Khoury, G. & Kelly, T. J., Jr. (1978) *J. Virol.* **25**, 193–201.

8. Padgett, B. L., Walker, D. L., RuRhein, G. M., Eckroede, R. J. & Dessel, R. (1971) *Lancet* **i**, 1257–1260.
9. Takemoto, K. K. & Mullarkey, M. F. (1973) *J. Virol.* **12**, 625–631.
10. Mullarkey, O. H., Hruska, J. F. & Takemoto, K. K. (1974) *J. Virol.* **13**, 1014–1019.
11. Mason, D. A. & Takemoto, K. K. (1976) *J. Virol.* **17**, 1060–1062.
12. Prives, C., Gilboa, E., Revel, M. & Winocour, E. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 457–461.
13. Crawford, L. V., Cole, C. N., Smith, A. E., Paucha, E., Tegtmeyer, P., Rundell, K. & Berg, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 117–121.
14. Paucha, E., Mellor, A., Harvey, R., Smith, A., Hewick, R. & Waterfield, M. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2165–2169.
15. Simmons, D. T. & Martin, M. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1131–1135.
16. Berk, A. J. & Sharp, P. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1274–1278.
17. Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200**, 494–502.
18. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Vockaert, G. & Ysebaert, M. (1978) *Nature (London)* **273**, 113–120.
19. Kelly, T. K. & Nathans, D. (1977) *Adv. Virus Res.* **21**, 85–173.
20. Sleigh, M. J., Topp, W. C., Hanich, R. & Sambrook, J. P. (1978) *Cell* **14**, 79–88.
21. Bouck, N., Beales, N., Shenk, T., Berg, P. & di Mayorca, G. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2473–2477.
22. Dhar, R., Lai, C.-J. & Khoury, G. (1978) *Cell* **13**, 345–358.
23. Howley, P. M., Khoury, G., Byrne, J. C., Takemoto, K. K. & Martin, M. A. (1975) *J. Virol.* **16**, 959–973.
24. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
25. Maniatis, T., Jeffrey, A. & Kleid, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 1184–1188.
26. Thimmappaya, B. & Weissman, S. M. (1977) *Cell* **11**, 837–843.
27. Ghosh, P. K., Reddy, B., Swinscoe, J., Choudhary, P. V., Lebowitz, P. & Weissman, S. M. (1978) *J. Biol. Chem.* **253**, 3643–3647.
28. Efstratiadis, A., Kafatos, F. P. & Maniatis, T. (1977) *Cell* **10**, 571–586.
29. Marotta, C. A., Wilson, J. T., Forget, B. G. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 5040–5051.