JOURNAL OF VIROLOGY, Feb. 1980, p. 619-630
0022-538X/80/02-0619/12$02.00/0

Vol. 33, No. 2

Polyoma Virus DNA: Complete Nucleotide Sequence of the Gene Which Codes for Polyoma Virus Capsid Protein VP1 and Overlaps the VP2/VP3 Genes†

EIICHI SOEDA,‡ JOHN R. ARRAND, AND BEVERLY E. GRIFFIN*
Imperial Cancer Research Fund, London WC2A 3PX, England

The nucleotide sequence of part of the late region of the polyoma virus genome was determined. It contains coding information for the major capsid protein VP1 and the C-terminal region of the minor proteins VP2 and VP3. In the sequence with the same polarity as late mRNA's, all coding frames are blocked by termination codons in a region around 48 units on the physical map. This is the region where the N-terminus of VP1 and the C-termini of VP2 and VP3 have been located (T. Hunter and W. Gibson, J. Virol. **28**:240-253, 1978; S. G. Siddell and A. E. Smith, J. Virol. **27**:427-431, 1978; Smith et al., Cell **9**:481-487, 1976). There are two long uninterrupted coding frames in the late region of polyoma virus DNA. One lies at the 5′ end of the sequence and contains potential coding sequences for VP2 and VP3. The other contains 383 consecutive sense codons starting with the ATG at nucleotide position 1,218, extends from 47.5 to 25.8 units counterclockwise on the physical map, and is located where the VP1 gene has been mapped. The VP1 gene overlaps the genes for proteins VP2/VP3 by 32 nucleotides and uses a different coding frame. From the DNA sequence, the amino acid sequence of VP1 was predicted. The proposed VP1 sequence is in good agreement with other data, namely, with the partial N-terminal amino acid sequence and the total amino acid composition. The VP1 coding frame terminates with a TAA codon at 25.8 map units. This is followed by an AATAAA sequence, which may act as a processing signal for the viral late mRNA's. When both nucleotide and amino acid sequences are compared with their counterparts in the related simian virus 40, extensive homologies are found over the entire region of the two viral genomes. Maximum homology appears to occur in those regions which code for the C-termini of the VP1 proteins. The overlap region of VP1 with VP2/VP3 of polyoma virus is shorter by 90 nucleotides than is that of simian virus 40 and shows very limited homology with the simian virus 40 sequence. This leads to the suggestion that the overlap segments of both viruses have been freed from stringency imposed on drifting during evolution and that proteins VP2 and VP3 of polyoma virus may have been truncated by the appearance of a termination codon within the sequence.

The genomes of polyoma virus (Py) and the related simian virus 40 (SV40) are packaged into 72 morphological units (hexons and pentons), which are arranged on irregular icosahedral surface lattices with T (triangulation number) = 7 (dextro) (2, 37). On the basis of stoichiometric measurements, Ester et al. (12) suggested that VP1 is represented a total of 360 times in each viral particle and that six molecules of VP1 make up each hexon. Christiansen et al. (9) have shown that 7S particles isolated from SV40 virions are identical to hexons, contain VP1, and are virtually free of VP2 and VP3. The location of VP2 and VP3 on the virion is not known. The available stoichiometric data (12) support the notion that the pentons are composed of 10 molecules of VP2 and VP3.

The genome of Py is comprised of a supercoiled double-stranded DNA of approximately 5,290 nucleotide pairs, which can be divided functionally into early and late genes. During the late lytic cycle of Py, after commencement of viral DNA replication, three viral RNAs appear in the cytoplasm of cells. These have sedimentation coefficients of 16S, 18S, and 19S (1, 7, 35), and they direct in vitro synthesis of the proteins VP1, VP3, and VP2, respectively (32,

† This paper and the accompanying papers are dedicated to M. G. P. Stoker on the occasion of his retirement as Director of Research of the Imperial Cancer Research Fund.

‡ On leave from the National Institute of Genetics, Misima 411, Japan.

51). The main body of 16S mRNA appears to be located from 48 to 26 map units counterclockwise on the physical map of Py DNA (22). The bodies of the 18S and 19S mRNA's span completely the region of 16S mRNA, starting from the proximal part of the late region.

Genetic data provide evidence that VP2 and VP3 are encoded within the proximal part of the late region, whereas VP1 is encoded within the distal part of the late region DNA (20, 39). Considering the sizes of VP2 and VP3, the C-termini of both proteins could be located around 48 map units. All tryptic peptides found in VP3 are also found in VP2, and the latter also has additional unique peptides. The peptides of VP1 are, however, different from those of VP2 and VP3 (19, 27). All lines of evidence suggest that the N-terminus of VP1 and the C-termini of VP2 and VP3 are located around 48 map units. By analogy with SV40 (10), the coding regions of these proteins may overlap each other, with different coding frames being used. The evidence is consistent with the topographical model of the late region of the physical map (Fig. 1), which shows positions of late mRNA's and capsid proteins. SV40 and Py have very similar growth cycles and overall genetic organizations, which suggests that the viruses might have diverged from a common origin (16). Heteroduplex analysis of the DNAs of the two viruses has revealed a limited nucleotide sequence homology, which is located within the region of the genomes in which the virion capsid proteins are coded (13). The strongest sequence homology occurred in the region which should code for the N-termini of VP1 in both Py and SV40 (16). Correspondingly, antisera raised against the virion proteins of SV40 were found to cross-react immunologically with the major capsid protein (VP1) of Py (50). In considering the evolution of genes, the speculation has been put forward that overlapping genes which use different coding frames should impose severe restrictions on genetic drift and that, consequently, the sequences in such regions should be highly conserved (40).

Extrapolation of these data and evolutionary arguments lead to the hypothesis that at least some parts of the VP1 genes of Py and SV40 should have conserved sequences. This prompted us to determine the nucleotide sequence of the Py VP1 gene and compare it with that of SV40 to examine gene conservation and to seek the relics of the ancestry of these viruses on the molecular level.

In this paper we report the complete nucleotide sequence of the VP1 coding region of Py and show the strong sequence homology with the VP1 coding region of SV40. Contrary to expectation, the overlap region of the genes coding for VP2/VP3 and VP1 is shorter (by 90 nucleotides) than in SV40, and the sequence homology in this region is small. These results suggest that the overlap regions of both viruses have not been subjected to severe constraints on genetic drift through evolution. Divergence of VP1 gene sequences is discussed from the aspect of molecular evolution.

MATERIALS AND METHODS

All procedures for propagation of Py strain A2 and isolation of the DNA have been described elsewhere (22).

Digestion of Py DNA with restriction enzymes. The restriction endonucleases used for cleaving Py DNA before labeling of the digestion products at the 3′ termini were *Hin*dIII, *Alu*I, *Hin*fI, *Hae*III, *Hpa*II, *Ava*II, and *Mbo*I. In addition to the above enzymes, *Mbo*II, *Pst*I, and *Hin*dII were used for the second cleavage of labeled fragments in order to obtain species labeled at only one end for sequencing. The conditions for enzyme digestions were as described previously (21, 22).

Terminal labeling of the fragments at the 3′

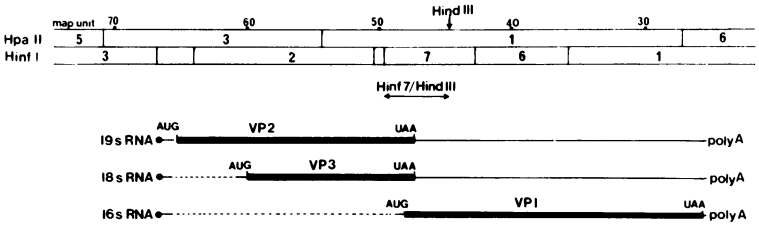

FIG. 1. *Topographical model of the Py late genes. The partial cleavage maps of the restriction endonucleases HpaII (22) and HinfI (23) are shown at the top on the linear late region of Py DNA. The DNA was divided into 100 map units, taking the single cleavage site of the restriction endonuclease EcoRI as a zero reference point (22). The three known viral late mRNA's (with polyadenylated tails) and their locations in terms of the genomes are shown. The coding regions of the three virion proteins (heavy solid lines) are shown, along with the regions spliced out of the mature late mRNA's (dotted lines) (30). The overlap region, which contains the sequence coding for the C-terminus of VP2/VP3 and the N-terminus of VP1, lies within the HindIII subfragment of HinfI-7. The sequence from this region is shown in Fig. 3.*

ends with T4-induced DNA polymerase. For more effective incorporation of ^{32}P and homogeneous labeling at the 3' ends of the restriction fragments, the method described previously (53, 54) was modified as follows. Restriction fragments with either blunt or 5' protruding ends, derived from 15 μg of form I DNA, were dissolved in 0.1-ml solutions containing 0.02 M Tris-hydrochloride buffer (pH 8.2), 0.007 M $MgCl_2$, and 0.002 M dithioerythritol and then labeled with T4-induced DNA polymerase (generously provided by N. Smolar) in the presence of 100 μCi of the α-P^{32}-labeled deoxyribonucleotide triphosphate (250 Ci/mmol; Radiochemical Centre, Amersham, England) complementary to the nucleotide residue at the 5' termini of the fragments and the other three cold deoxyribonucleoside triphosphates (each at 50 μM). Incubation was performed at 15°C for labeling fragments with protruding 5' ends and at 25°C for fragments with blunt ends. The incorporation of ^{32}P into the fragments was followed by thin-layer chromatography on a polyethyleneimine plate (Schleicher & Schuell) developed with 0.75 M potassium phosphate buffer, pH 3.5.

Fractionation of the labeled fragments on a polyacrylamide gel and elution from the gel. The labeled fragments were loaded onto 4 to 6% polyacrylamide slab gels (0.3 by 20 by 40 cm) and electrophoresed overnight at 220 V, using Tris-hydrochloride (pH 7.8)-EDTA buffer (26). The separated bands in the gel were visualized on X-ray film (Fuji medical RX) after autoradiography for about 10 min. The labeled fragments were excised and eluted electrophoretically from the gel into dialysis bags (17), and the DNA isolated in this way was concentrated by ethanol precipitation.

Preparation of fragments labeled at only one end. The isolated fragments were cleaved with a second restriction enzyme and applied to a 6% polyacrylamide slab gel, and individual fragments were again separated by electrophoresis. These fragments, labeled at one 3' end only, were eluted after the gel was crushed. They were isolated by standard procedures and then sequenced by the method described by Maxam and Gilbert (38).

Sequencing gel. Polyacrylamide denaturing gels (12%; 0.6 by 300 by 900 mm) were used for sequencing. When the blue dye marker (xylene cyanol) had migrated one-fifth of the way from the bottom of the gel in the first electrophoretic separation, a second portion of the sample was usually applied to the gel. After further electrophoresis, the gel was wrapped with a thin plastic film and exposed to X-ray film for 2 to 4 days at −70°C, using intensifying screens (mach-2; Fuji).

RESULTS

The body of the late 16S RNA, which is found in Py-infected cells and codes for VP1, lies from about 48 to 26 units counterclockwise on the physical map of Py DNA (35). This region is contained within the *Hpa*II-1 and -6 restriction endonuclease fragments and spans the *Hae*III-19, -16, -4, -10, -22, -11, -20, and -6 fragments (counterclockwise) (Fig. 2) (21). The nucleotide sequence in this region was deduced by the chemical method of Maxam and Gilbert (38). For this purpose, (i) restriction fragments with either blunt or protruding ends were labeled specifically at their 3' ends with the T4-induced DNA polymerase in the presence of three cold deoxyribonucleoside triphosphates and one ^{32}P-labeled deoxyribonucleoside triphosphate (see above); (ii) long thin sequencing gels were used from which a sequence of more than 200 nucleotides could be deduced unambiguously from one labeled fragment and a further 50 or so nucleotides could be tentatively deduced; and (iii) to prevent errors, most sequences were determined more than once from different strands of the DNA and from different restriction endonuclease fragments.

The sequence extending from 49.0 to 25.2 map units could be nearly completely deduced from autoradiograms of gels similar to those shown in Fig. 3. The VP1 coding region should lie within this sequence. Where anomalous spaces between

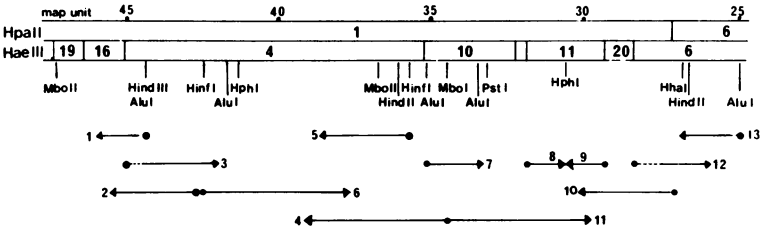

FIG. 2. *Labeled fragments used in nucleotide sequencing. The DNA coding for the VP1 protein was sequenced by the method of Maxam and Gilbert (38), using restriction fragments labeled at one 3' end only (solid circles). From those points, more than 200 nucleotides could be sequenced for each fragment in the directions indicated by the arrows. Numbers on the arrows are arbitrary and are our own reference numbers. Any ambiguity observed in a sequence adjacent to a labeled end was corrected by sequencing over the same area from other restriction endonuclease sites. The sequence was also confirmed by other methods (see text) (data not shown). Cleavage sites of the restriction enzymes have been reported elsewhere (21-23).*

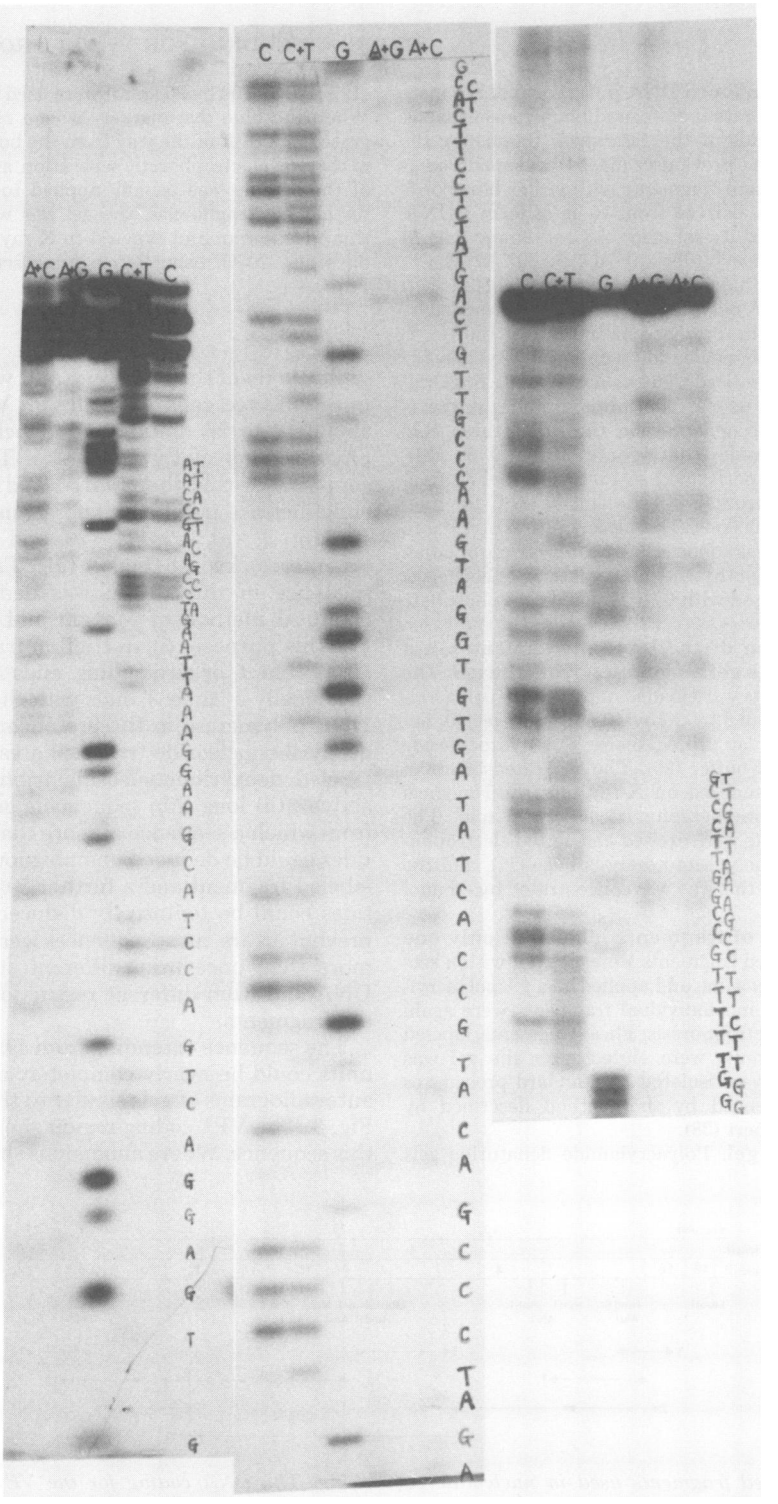

FIG. 3. *Autoradiograms of sequencing gels. Restriction fragments with either blunt or 5′ protruding ends were treated with T4-induced DNA polymerase in the presence of four deoxyribonucleoside triphosphates, one of which was labeled at the α position with* ^{32}P. *Fragments labeled at one 3′ end only (see text) were submitted to partial base-specific chemical degradation (38) and separated electrophoretically on 12 or 20% polyacrylamide gels (0.6 by 300 by 900 mm) in the presence of 7 M urea and Tris-borate buffer, pH 8.2. After electrophoresis, gels were exposed to X-ray film. Only gels with sequence corresponding to the N-terminal position of VP1 are shown. (Other gels have been presented for substantiation.) They contain a sequence which corresponds to nucleotides between positions 1,138 and 1,270 and includes the VP2/VP3 and VP1 overlapping sequences (see Fig. 1 and 4).*

nucleotides were observed on sequencing gels (most commonly seen in cytidine-specific tracks), the sequence of this particular region of the DNA was determined from the opposite strand. Thus, the 1,275-nucleotide sequence was determined by joining the overlapping sequences from different restriction fragments with the same polarity and (complementary) sequences of different fragments with opposite polarity. Corroborative evidence also comes from qualitative pyrimidine tract analysis (21) and DNA sequencing in which the dideoxy method developed by Sanger et al. (48) was used (data not shown). When the sequence having the same polarity as late mRNA's is numbered from the midpoint of the *Hpa*II restriction endonuclease site between fragments *Hpa*II-3 and -5 (conventionally defined as the boundary between the early and late regions), the present sequence is located from nucleotide 1,133 to nucleotide 2,407 in the late region (Fig. 4). Several restriction enzyme sites are predicted within this region, the positions of which are more or less consistent with sites determined by other methods (21-23). In Fig. 4, the termination codons found in each of the three potential coding frames (designated 1, 2, and 3) are indicated. In the sequence analysis of DNA from this region, all three coding frames were found to contain termination codons at about 48 map units (or the region around nucleotide 1,240). This suggests that no protein can be read beyond this region without processing of the mRNA, or alternatively, that both the termination site of VP2/VP3 and the initiation site of VP1 might reside in this region, different coding frames being used in the two cases. DNA sequence and protein chemistry data agree with the latter hypothesis (see below). It has been found that the coding frame (designated frame 1) which contains the ATG initiation codon for VP2 extends for 966 nucleotides before a termination codon is reached (that is, from late region nucleotide 290 to nucleotide 1,246) (3). The body of the 16S VP2 mRNA has been mapped within this region (52). Therefore, as many termination codons appear in the other two frames, this coding frame can be assigned with some confidence as that used for all (or at least most) of VP2 and VP3. (There is as yet no amino acid sequence data on the C-termini of these proteins.) Another coding frame (designated frame 2) contains no termination codons between nucleotides 1,211 and 2,375 (see below). This frame overlaps the potential coding frame for VP2/VP3 by 35 nucleotides. It contains an ATG at nucleotide position 1,218 which appears to be the initiation codon of VP1.

From the DNA sequence, the amino acid sequence of VP1 can be predicted. Recently, Hewick et al. (28) determined the partial amino-terminal sequence of a VP1 protein synthesized in vitro. Their data, derived from the first 20 cycles of the sequence, completely match that predicted from the DNA sequence (Fig. 5). They detected no methionine in their products, so possibly the initiating methionine was lost from the amino terminus of the protein during in vitro synthesis.

The results described above indicate that the VP2/VP3 genes overlap the VP1 gene, with different coding frames being used for the viral late proteins. The actual coding sequence for VP2/VP3 extends 32 nucleotides beyond the

```
                                    -GAGTCAC TCCTGACTGG ATGCTTCCTT TAATTCTAGG GCTGTACGGT GATATCACAC CTACTTGGGC   1200
AACAGTCATA GAGGAAGATG GCCCCCAAAA GAAAAAGCGG CGTCTCTAAA TGCGAGACAA AATGTACAAA GGCCTGTCCA AGACCCGCAC CCGTTCCCAA   1300
ACTGCTTATT AAAGGGGGTA TGGAGGTGCT GGACCTTGTG ACAGGGCCAG ACAGTGTGAC AGAAATAGAA GCTTTTCTGA ACCCCAGAAT GGGGCAGCCA   1400
CCCACCCCTG AAAGCCTAAC AGAGGGAGGG CAATACTATG GTTGGAGCAG AGGGATTAAT TTGGCTACAT CAGATACGTG GATTCCCCGA AATAATACAC   1500
TTCCCACATG GAGTATGGCA AAGTCCAGCT TCCCATGCCT CAATGAGGAC CTCACCTGTG ACACCCTACA AATGTGGGAG GCAGTCTCAG TGAAAACCGA   1600
GGTGGTGGGC TCTGGCTCAC TGTTAGATGT GCATGGGTTC AACAAAACCC ACAGATTCAG TAAACACAAA GGAAATTCCA CTCCAGTGGA AGGCAGCCAA   1700
TATCATGTGT TTGCTGGTGG CGGGGAACCG CTTGACCTCC AGGGACTTGT GACAGATGCC AGAACAAAAT ACAAGGAAGA AGGGGTAGTA ACAATCAAAA   1800
CAATCACAAA GAAGGACATG GTCAACAAAG ACCAAGTCCT GAATCCAATT AGCAAGGCCA AGCTGGATAA GGACGGAATG TATCCAGTTG AAATCTGGCA   1900
TCCAGATCCA GCAAAAAATG AGAACACAAG GTACTTTGGC AATTACACTG GAGGCACAAC GACTCCACCC GTCCTGCAGT TCACAAACAC CCTGACAACT   2000
GTGCTCCTAG ATGAAAATGG AGTTGGGCCC CTCTGTAAAG GAGAGGGCCT ATACCTCTCC TGTGTAGATA TAATGGGCTG GAGAGTTACA AGAAACTATG   2100
TGTCATCACT GGAGAAGGGC TTTCCCAGAT ATTTCAAAAT CACCCTGAGA AAAAGATGGG TCAAAAATCC CTATCCCATG GCCTCCCTCA TAAGTTCCCT   2200
TTTCAACAAC ATGCTCCCCC AAGTGCAGGG CCAACCCATG GAAGGGGAGA ACACCCAGGT AGAGGAGGTT AGAGTGTATG ATGGGACTGA ACCTGTACCG   2300
GGGGACCCTG ATATGACGCG CTATGTTGAC CGCTTTGGAA AAACAAAGAC TGTATTTCCT GGAAATTAAT GTTTATTCAA TAAACTGTGT ATTCAGCT     2400
```

FIG. 4. *Total nucleotide sequence of the VP1 coding region. The total nucleotide sequence of the gene which codes for VP1 and the sequence which precedes the initiation codon at position 1,218 and the termination codon at position 2,367 are shown. The DNA sequence was constructed by joining overlapping regions of the sequenced fragments. Termination codons are indicated by brackets and numbers are given to each potential coding frame. Nucleotides are numbered from the midpoint of the HpaII site at the junction of fragments HpaII-3 and -5, according to the convention adopted previously (3, 53).*

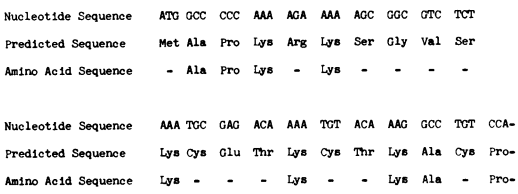

FIG. 5. *Coincidence of the partial amino acid sequence from the N-terminus of VP1 as determined by protein chemistry compared with the amino acid sequence predicted from the nucleotide sequence. The second coding frame beginning at nucleotide 1,218 (Fig. 4) allows an amino acid sequence to be predicted for VP1, using the accepted triplet codons. The partial sequence of the N-terminus of VP1 as determined by protein chemistry (28) matches the predicted one for the first 20 cycles in the amino acid sequence analysis.*

VP1 initiation site and terminates with the TAA at nucleotide 1,247, which lies within the VP1 coding region.

DISCUSSION

Assignment of the VP1 coding region on the DNA sequence. From the ATG triplet at nucleotide 1,218, which appears to be the initiation codon of VP1, the same coding frame remains open for 1,148 nucleotides, that is, for 383 consecutive triplet codons. This sequence extends from about 47.5 to 25.7 units on the physical map of Py DNA (22). The other two potential coding frames are blocked by a number of termination codons within each frame; the longest sequence without a termination codon is 234 nucleotides long. This, if used, would code for a polypeptide of only 78 amino acids. Therefore, the 1,149-nucleotide sequence can be assumed to code for VP1, and the amino acid sequence of this protein can be predicted according to the accepted triplet code. The amino acid composition of VP1 as purified from virions is known (14, 29) and is in good agreement with the amino acid composition predicted by DNA sequencing (Table 1). Fey and Hirt (14), for example, found 31 lysines, 18 or 19 arginines, and 12 methionines in VP1, which is in good agreement with our predicted values of 29, 16, and 12, respectively, and provides further evidence for the validity of our assignment. The molecular weight of VP1, as calculated from the predicted amino acid sequence, is 42,458, which is in contrast to the apparent molecular weights of 45,000 to 48,000 estimated from electrophoretic mobility on sodium dodecyl sulfate-polyacrylamide gels (14, 19, 27).

Code word usage and comparison with SV40. The predicted code words used for Py protein VP1 are shown in Fig. 6, where they are compared with those predicted for SV40 VP1 (41). The overall amino acid compositions of the two VP1 proteins are similar. The code words reflect the drastic deficiency in the use of the doublet CG in viral codons, which is in agreement with the relative infrequency of this doublet in the corresponding host cell DNAs (46). (In some eucaryotic systems, this dinucleotide is protected by methylation [4, 11].) Analysis of the Py DNA sequence shows that the CG doublet appears with the frequency expected for a random event between the third and the first positions at the junction of consecutive codons, however. The other relative code word biases pointed out in SV40 (15, 44) do not appear to be applicable to Py VP1, despite the similar amino acid compositions of the two proteins.

TABLE 1. *Amino acid composition of Py VP1 protein and the primary products of VP1 predicted from the nucleotide sequence*[a]

Amino acid	Amino acid composition of VP1 (no. of residues)	
	DNA sequence	Protein
Cysteine	7	5.6
Aspartic acid	19	45.1
Asparagine	20	
Glutamic acid	23	37.2
Glutamine	11	
Methionine	12	9.8
Threonine	36	36.5
Serine	22	16.2
Glycine	35	39.9
Alanine	12	13.2
Valine	32	35.0
Leucine	30	32.3
Isoleucine	11	10.9
Proline	31	28.6
Histidine	5	5.3
Lysine	29	27.4
Arginine	16	16.5
Tyrosine	13	13.2
Phenylalanine	12	9.0
Tryptophan	7	

[a] The composition of the VP1 protein was recalculated from published data (29), assuming VP1 to be composed of 383 amino acid residues.

Potential recognition sequences for ribosome binding and splicing of mRNA. The region preceding the VP1 initiation site is characterized by a long inverted repeat sequence, which possibly forms a hairpin loop in the mRNA (Fig. 7). Such a structure might be utilized in the recognition events involved in the processing which leads to mature 16S mRNA or might serve a negative function by preventing aberrant binding of ribosomes at internal sites of 18S and 19S RNAs (see below). The structural features of eucaryotic mRNA's that specify the

	U		C		A		G		
U	Phe	6 (10)	Ser	2 (5)	Tyr	8 (6)	Cys	5 (3)	U
		6 (1)		5 (2)		5 (5)		2 (4)	C
	Leu	1 (8)		5 (1)	Ochre	1 0	Term	0 0	A
		1 (9)		0 (0)	Amber	0 0	Trp	7 (3)	G
C	Leu	6 (4)	Pro	4 (12)	His	3 (4)	Arg	0 (0)	U
		8 (1)		15 (5)		2 (2)		2 (0)	C
		4 (2)		10 (11)	Gln	6 (8)		1 (0)	A
		10 (4)		2 (0)		5 (10)		0 (1)	G
A	Ile	4 (12)	Thr	6 (12)	Asn	11 (11)	Ser	4 (6)	U
		4 (0)		8 (9)		9 (8)		6 (3)	C
		3 (3)		19 (11)	Lys	19 (19)	Arg	12 (7)	A
	Met	12 (8)		3 (0)		10 (6)		1 (5)	G
G	Val	6 (15)	Ala	3 (13)	Asp	9 (9)	Gly	3 (5)	U
		6 (1)		5 (1)		10 (11)		11 (2)	C
		6 (4)		4 (4)	Glu	11 (11)		9 (12)	A
		14 (10)		0 (0)		12 (13)		12 (10)	G

FIG. 6. *Codon usage of VP1 as predicted from the DNA sequence of Py. For comparison, the codon usages found in the gene encoding VP1 in SV40 (41) are given in parentheses.*

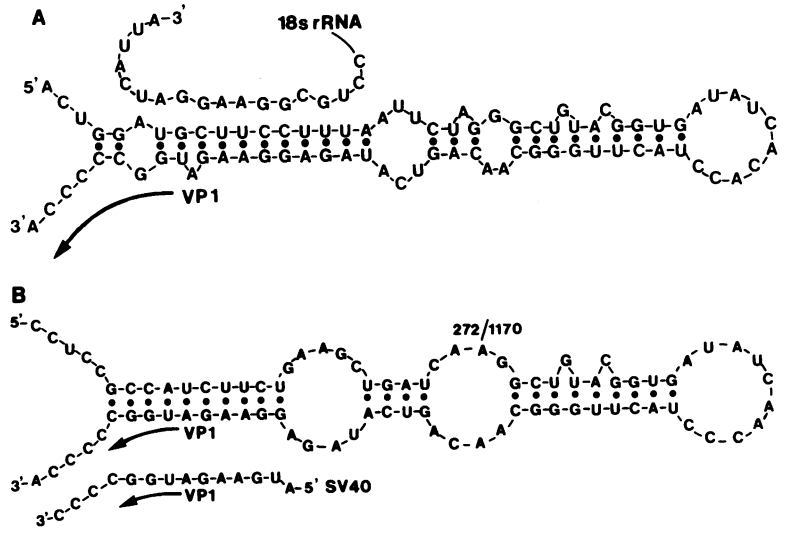

FIG. 7. *Putative structural features of the nucleotide sequence which precedes the initiation site for translation of VP1. (A) The region preceding the VP1 coding region of Py is characterized by a long inverted repeat sequence; the same is true for SV40 (45). This may form a hairpin structure at the cRNA level, as shown. The consensus nucleotide sequence observed in eucaryotic 18S rRNA (25) and postulated to have a role in ribosome-mRNA interactions has a sequence which can potentially pair with the late cRNA in this region, in a manner such as that indicated. Whether this is functional or fortuitous is not known. The AUG initiation codon for VP1 appears as indicated toward the 3' end of the sequence. (B) The region preceding the VP1 coding region of Py in mature 16S mRNA is characterized by a long inverted repeat sequence, similar to that shown in (A). It is postulated (see text) that 16S mRNA involves splicing between the leader sequence from a position which corresponds to a nucleotide found in the DNA at position 272 (3) and the messenger body sequence from position 1,170, as illustrated. The leader sequence TCTTCT could, at least theoretically, pair with the consensus nucleotide sequence in eucaryotic 18S rRNA or with the VP1 initiation codon, as illustrated here. The sequence found near the 5' end of the SV40 VP1 mRNA is similar to that found in Py. It is interesting that it is the second of the two potential AUG initiation codons (45) in SV40 which lies in a position equivalent to the Py initiation codon. Attempts to suggest functional roles for structures such as those shown here are complicated by the factors discussed in the text. If transcription and translational controls in Py are analogous to those in SV40, assigning a role to the leader sequence becomes further complicated by the fact that virtually all of it can be deleted without affecting the viability of the latter virus (55). There is as yet no evidence, however, to say that the two viruses should be considered in parallel in this respect.*

initiation of translation remain obscure, although a number of the nucleotide sequences preceding the initiation sites for translation have been determined. From data which showed that a purine-rich sequence (5'-UGCGGAAGGAU-3') near the 3' termini of 18S ribosomal RNAs is common to mice, silkworms, and wheat, Hagenbüchle et al. (25) proposed that this sequence might be involved in mRNA–18S rRNA interactions. In this context, it is noteworthy that the DNA sequence from nucleotides 1,150 to 1,161 (5'-ATGCTTCCTTT-3'), which could pair with the 18S rRNA sequence, is buried in a hairpin structure such as is shown in Fig. 7A. It is interesting to note that the sequence TCTTCTG, which occurs within the noncoding part of the genome and corresponds to reiterated sequences found in the leaders of late mRNA's (nucleotides 255 to 261) (3), can also pair with this segment of 18S rRNA. (This sequence could also be buried in a hairpin loop similar to that shown in Fig. 7A [see below].) Because the leader (untranslated) sequences of Py late 16S mRNA appear to be ligated to the body of the messenger by a sequence from around this region (see below and Fig. 7B), this may suggest that leader sequences themselves provide additional features necessary for efficient recognition of enzymes involved in initiation of translation or for ribosomal binding.

As mentioned above, the 16S late mRNA which codes for VP1 is a product which is complementary to a discontinuous region of the late gene and has presumably arisen from a precursor RNA by some splicing mechanism (30, 34). The 5' end of the body of 16S mRNA is joined at about 48 map units to a leader region from around 67 map units (Fig. 1). From a consideration of splicing rules (6, 8, 18, 49) and data on 16S mRNA obtained by S1 mapping procedures (33), the intervening DNA sequence (that is the sequence absent in the messenger) should come between nucleotides lying at positions 273 (3) and 1,169 (Fig. 4). If this is the case, the sequences which specify the splice for the VP1 mRNA are as follows:

```
                273
5'-ATCAA/GTAAGT -------
                                1169
                      -----TTCTAG/GGCTAG-3'
```

In such a case the body of the messenger would contain 48 nucleotides that are noncoding. In the corresponding SV40 mRNA, 42 nucleotides from the body of the messenger precede the initiation codon for VP1, and the corresponding splicing sequences have been found to be 5'-AACTG/GTAAGT and TTCTAG/GCCTGT-3', respectively (18). These are remarkably similar to those postulated for Py.

If the splicing postulate that leads to mature 16S mRNA is correct, a hairpin model similar to that proposed in Fig. 7A can be formed, which involves base pairing between sequences from the leader region and sequences around the initiation codon of VP1 (Fig. 7B). This model also has putative ribosomal binding sequences buried within a hairpin structure. The fact that such structure could be formed in the mature messenger species, as well as in the presumed precursor to mature messenger, makes it difficult to suggest roles for the structures. Presumably, an important role for secondary structures in a precursor to mRNA would be to ensure correct splicing. Structures leading to correct splicing would presumably be irrelevant in a mature messenger. Since so little is known about the mechanisms of transcription in eucaryotic systems, splicing, etc., it may be premature at this stage to do other than point out the existence of such interesting secondary structures and suggest that they may have a biological function which at present is not obvious. As illustrated in Fig. 7B, the sequence at the 5' end of the SV40 messenger for VP1 is similar to that seen for the corresponding Py messenger. If a direct correlation between the two messengers is relevant, the second (rather than the first) ATG should be the initiation codon for VP1 in SV40.

Potential termination sequence for VP1 and the structural features around the polyadenylate ends of the late mRNA's. The region from nucleotides 2,370 to 2,385 is characterized by the presence of an inverted repeat sequence which has a twofold rotational axis of symmetry and contains the sequence AATAAA in both strands. The potential open frame (frame 1) coding for VP1 terminates at nucleotide 2,367 with a TAA which precedes this region by one nucleotide. (The frame presumed to code for large T antigen, a protein essential for viral DNA replication, terminates at nucleotide 2,379 with TGA on the opposite [complementary] strand in this region [52a].) The large T1 RNase oligoribonucleotides of polyadenylated late mRNA's have been isolated and sequenced (S. Legon, personal communication). The heptanucleotide UAUUCAG, which was predicted from the DNA sequence (Fig. 4, between nucleotides 2,390 and 2,396), was not found in the mRNA's, whereas all of the other T1 oligoribonucleotides predicted from the DNA sequence were observed. This indicates that the polyadenylated tails of the late mRNA's probably start somewhere between nucleotides 2,390 and 2,396. In agreement with this, many reports have accumulated that the hexanucleotide AAUAAA is

present in the noncoding regions of eucaryotic mRNA's and precedes polyadenylated tails (42, 43), suggesting that the sequence AAUAAA may be involved either in the processing of mRNA or in a polyadenylation step. Black et al. (5) have reported, however, that this sequence is absent from the noncoding regions at the 3′ ends of some picornavirus genome RNAs, although these RNAs are polyadenylated. Therefore, the sequence AAUAAA situated near the polyadenylated ends of mRNA's may be only one of the prerequisites for the processing of the mRNA's.

Sequence homologies between SV40 and Py VP1 genes. Since the complete nucleotide sequences of the VP1 genes of SV40 (15, 41, 45) and Py have been determined, it is possible to look for ancestral relationships at the molecular level. When both the nucleotide and amino acid sequences of the two viral species are aligned (inserting gaps to maximize homologies), extensive homologies are found throughout most of the VP1 genes (Fig. 8). (It is obvious, however, that there are also some unique regions in each of the two species, particularly around the region where the longest gap has been inserted [Fig. 8].) There are, for example, several cysteine residues in both virus proteins, which would allow them to adopt distinct conformations. The

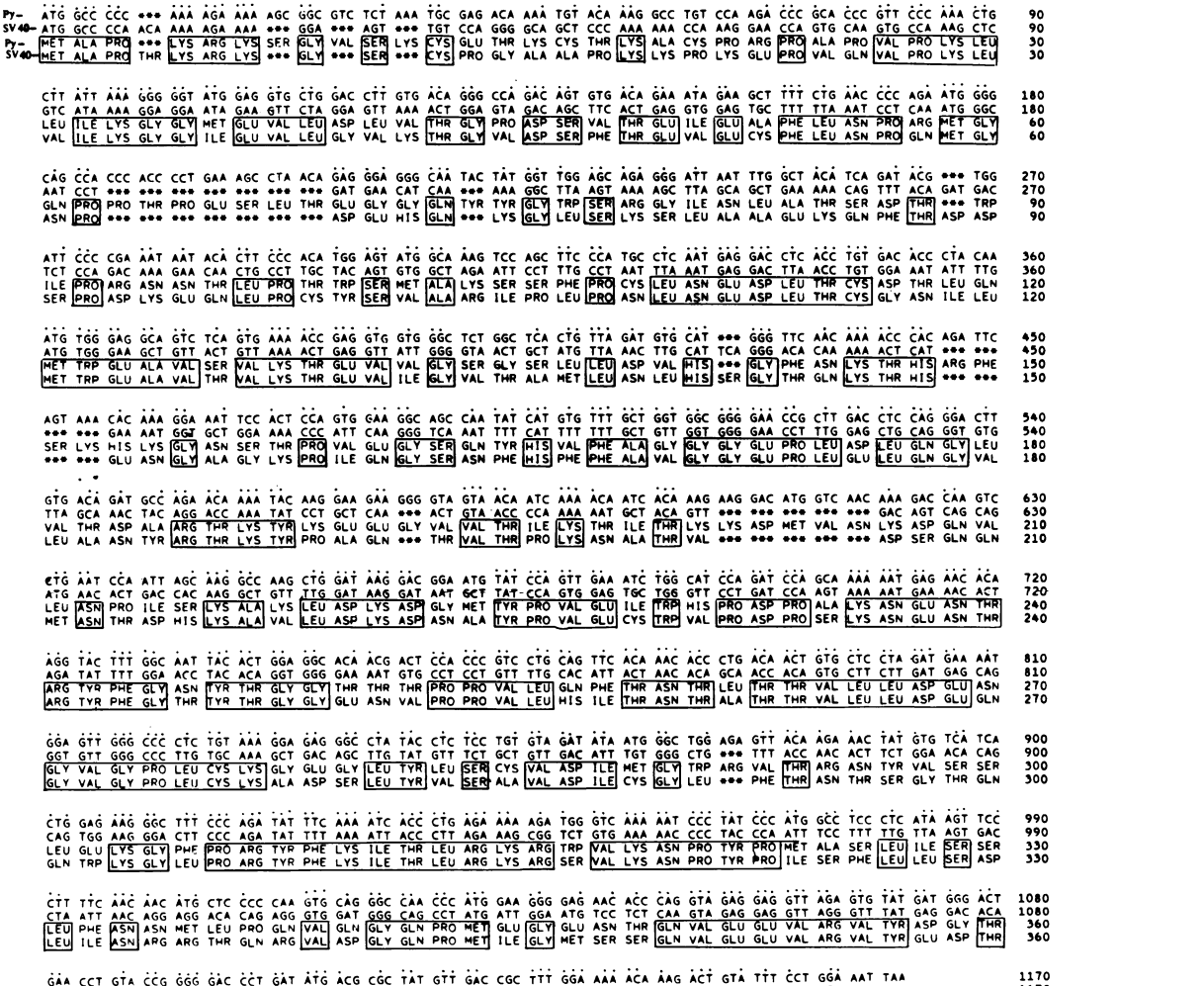

FIG. 8. *Comparison of nucleotide and amino acid sequences of the VP1 genes and gene products of Py and SV40. The VP1 sequences of Py and SV40 are aligned so as to maximize homology. Identical peptides and nucleotides at corresponding positions are indicated by the boxes and dots, respectively. The use of numbers in this figure is obvious and bears no relationship to those used either in the text or in Fig. 4.*

presence of these cysteines might permit the following observation to be explained. Walter and Deppert (56) showed that the hexons in the virus icosahedral shell are covalently bound to each other (but not to VP2 or VP3). Using sodium dodecyl sulfate in the absence of reducing agents, they isolated from intact Py viral particles large polymeric structures that sedimented at the same rate as shells of empty particles and by subsequent treatment were found to consist only of VP1 monomers. Because the addition of mercaptoethanol or dithiothreitol causes viral shells to dissolve, it seems likely that in Py, VP1 molecules are connected to each other by disulfide bridges.

The distribution of homologous peptides in the Py and SV40 VP1 sequences is not uniform along the peptide chains and may reflect responses to different selective constraints during evolution. Heteroduplex analysis with covalently linked full-length DNAs of SV40 and Py suggests that the strongest homologies are located around the N-termini of the VP1 genes (13). The present data, on the contrary, show that most of the larger homologous peptides are found in regions of the genes coding for the C-termini. This disparity may be a reflection of the two different methods of analysis. Peptide homologies are predicted from DNA sequences and from the genetic code, as determined for procaryotic systems. It is not clear at the moment what factors are most important in heteroduplex formation (31). Antisera raised against disrupted virions of SV40 can cross-react immunologically with VP1 of Py (50), suggesting that there are common antigenic determinants in at least parts of the VP1 regions of the two viruses. This intrinsic property of VP1 might reside in regions conserved between the two viruses.

Preservation of overlapping genes. Overlapping genes which use different coding frames were discovered in the ϕX174 genome (47). It was suggested subsequently that such genes must be highly preserved in genetic drift during evolution, because, once the overlap expression is established, all mutations in one gene could also alter the other gene (40). In SV40, the genes for VP2/VP3 overlap the VP1 gene by 122 nucleotides (10). This is the region where the strongest homology was found in the heteroduplex analysis of SV40 and Py DNAs (13). When the sequences of SV40 and Py DNAs are ordered in such a way as to align initiation sites of the VP1 genes, it has been found that the overlap region of Py is only 32 nucleotides long (shorter by 90 nucleotides than the overlap region of SV40). If the Py termination codons in the region corresponding to the remainder of the SV40 VP2/VP3 genes are ignored, still only limited homology is observed in these regions of the two genomes. These data suggest that the overlap regions of both viral species appear to be more or less free from the stringency of evolutionary drift. The C-termini of the VP2 and VP3 proteins of Py may have been truncated by a termination signal emerging within the coding frame, or conversely, the SV40 C-termini may have been lengthened by the loss of a termination signal.

Recently, two complete nucleotide sequences of the related bacteriophages G4 (24) and ϕX174 (47) have been determined. In the overlap regions, where different coding frames are used, the frequency of nucleotide change was found to be slightly lower than in the nonoverlap regions, thus permitting a number of amino acid alterations to occur between the relevant G4 and ϕX174 proteins (24). In general, it is known that the domains (parts of proteins with biological activity) are well conserved during evolution. The rest of a protein is less subject to stringency and more free to drift during evolution. It seems unlikely, therefore, that two domains, endowed with different functions, are encoded within the same regions in the overlapping genes in Py and SV40.

Divergence of the sequence. Mutations can occur in genes with the same function without involving amino acid changes. These mutations can lead to the divergence of the nucleotide sequences of the species. Assuming an ancestral relationship between Py and SV40, this type of change appears to have occurred at a higher rate than the mutations which produce changes in the amino acids themselves. The former type of change mainly involves alterations of nucleotides in the third positions of the codons (36). However, limited alterations of nucleotides at the first and the second positions, in addition to the third position, can also occur in codons for leucine, arginine and serine. Table 2 shows the

TABLE 2. *Nucleotide substitution patterns of leucine, arginine, and serine codons in synonymous parts of the SV40 and Py VP1 genes*[a]

Codon	No. of codons where no change occurs	No. of codons with change(s) at position(s):						
		1	2	3	1 and 2	1 and 3	2 and 3	1, 2, and 3
Leucine	1	2		8		8		
Arginine	3	0		3		2		
Serine	2	0	0	3	1	0	0	1

[a] Regions of SV40 and Py where homology appears to exist.

codon changes for these amino acids in synonymous regions of the DNAs. It is of interest to note that many of the nucleotide changes which occur without a change in amino acids are actually two-base changes. Moreover, in the case of leucine, which provides most examples, nucleotides appear to be replaced with an equal frequency at the first and third positions of the codons, suggesting that mutations have occurred selectively on genes during evolution.

Most mutations in the first and second positions of triplet nucleotides, however, lead to alterations of amino acids. In some cases, depending on the chemical nature of the amino acids and the functional constraints on the proteins, such changes may be without significant effect on gene function. A tendency for substitution of amino acids similar in chemical and structural properties can be observed in amino acids at the corresponding sites of the sequences in the viral proteins (Fig. 7). If these are taken into account, there is even greater homology between the SV40 and Py VP1 proteins than indicated in Fig. 7. If such substitutions are scored as neutral, it appears reasonable to postulate that VP1 serves the same function in both Py and SV40.

ACKNOWLEDGMENTS

We thank T. Maruyama for helpful discussion and computer programming of the sequences and M. D. Waterfield and R. Kamen for critical reading of the manuscript. We are grateful to S. Legon for unpublished data on late mRNA's, to R. Kamen for making available to us his data on late mRNA's before publication and his ideas on messenger processing, and to R. Hewick and A. Mellor for information received about the amino acid sequence of VP1 before publication. We are grateful to Jane Walsh for her excellent technical assistance.

LITERATURE CITED

1. **Acheson, N. H.** 1976. Transcription during productive infection with polyoma virus and simian virus 40. Cell **8**:1-12.
2. **Anderer, F. A., H. D. Schlumberger, M. A. Koch, H. Frank, and H. J. Eggers.** 1967. Structure of simian virus 40. II. Symmetry and components of the virus particle. J. Mol. Biol. **32**:511-523.
3. **Arrand, J. R., E. Soeda, J. E. Walsh, N. Smolar, and B. E. Griffin.** 1980. Polyoma virus DNA: sequence from the late region that specifies the leader sequence for late mRNA and codes for VP2, VP3, and the N-terminus of VP1. J. Virol. **33**:606-618.
4. **Bird, A. P., and E. M. Southern.** 1978. Use of restriction enzymes to study eukaryotic DNA methylation. I. Methylation pattern in ribosomal DNA from *Xenopus laevis*. J. Mol. Biol. **118**:27-47.
5. **Black, D. N., D. J. Rowlands, T. J. R. Harris, and F. Brown.** 1978. 3'-Terminal nucleotide sequences in the genome RNA of picornaviruses. Nature (London) **276**: 298-301.
6. **Breathnach, R., C. Benoist, K. O'Hare, F. Gannon, and P. Chambon.** 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. U.S.A. **75**:4853-4857.
7. **Buetti, E.** 1974. Characterization of late polyoma mRNA. J. Virol. **14**:249-260.
8. **Catteral, J. F., B. W. O'Malley, M. A. Robertson, R. Staden, Y. Tanaka, and G. G. Brownlee.** 1978. Nucleotide sequence homology at 12 intron-exon junctions in the chick ovalbumin gene. Nature (London) **275**: 510-513.
9. **Christiansen, G., T. Landers, J. Griffith, and P. Berg.** 1977. Characterization of components released by alkali disruption of simian virus 40. J. Virol. **21**:1079-1084.
10. **Contreras, R., R. Rogiers, A. Van de Voorde, and W. Fiers.** 1977. Overlapping of the VP2-VP3 gene and VP1 gene in the SV40 genome. Cell **12**:529-538.
11. **David, I. B., D. Brown, and R. H. Reeder.** 1970. Composition and structure of chromosomal and amplified ribosomal DNA's of *Xenopus laevis*. J. Mol. Biol. **51**: 341-360.
12. **Ester, M. K., E. S. Huang, and J. S. Pagano.** 1971. Structural polypeptides of simian virus 40. J. Virol. **7**: 635-641.
13. **Ferguson, J., and R. W. Davis.** 1975. An electron microscopic method for studying and mapping the region of weak sequence homology between simian virus 40 and polyoma DNAs. J. Mol. Biol. **94**:135-149.
14. **Fey, G., and B. Hirt.** 1975. Fingerprints of polyoma virus proteins and mouse histones. Cold Spring Harbor Symp. Quant. Biol. **39**:235-241.
15. **Fiers, W., R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, and M. Ysebaert.** 1978. Complete nucleotide sequence of SV40 DNA. Nature (London) **273**:113-120.
16. **Fried, M., and B. E. Griffin.** 1977. Organization of the genomes of polyoma virus and SV40. Adv. Cancer Res. **24**:67-113.
17. **Galibert, F., J. Sedat, and E. Ziff.** 1974. Direct determination of DNA nucleotide sequences: structure of a fragment of bacteriophage ϕX174 DNA. J. Mol. Biol. **87**:377-407.
18. **Ghosh, P. K., V. B. Reddy, J. Swinscoe, P. V. Choudary, P. Lebowitz, and S. M. Weissman.** 1978. The 5'-terminal leader sequence of late 16S mRNA from cells infected with simian virus 40. J. Biol. Chem. **253**: 3643-3647.
19. **Gibson, W.** 1974. Polyoma virus proteins: a description of the structural proteins of the virion based on polyacrylamide gel electrophoresis and peptide analysis. Virology **62**:319-336.
20. **Gibson, W., T. Hunter, B. Cogen, and W. Eckhart.** 1977. Altered virion proteins of a temperature-sensitive mutant of polyoma virus ts59. Virology **80**:21-41.
21. **Griffin, B. E.** 1977. Fine structure of polyoma virus DNA. J. Mol. Biol. **117**:447-471.
22. **Griffin, B. E., M. Fried, and A. Cowie.** 1974. Polyoma DNA: a physical map. Proc. Natl. Acad. Sci. U.S.A. **71**: 2077-2081.
23. **Griffin, B. E., and C. Maddock.** 1979. New classes of viable deletion mutants in the early region of polyoma virus. J. Virol. **31**:645-656.
24. **Godson, G. N., B. G. Barrell, R. Staden, and J. C. Fiddes.** 1978. Nucleotide sequence of bacteriophage G4 DNA. Nature (London) **276**:336-347.
25. **Hagenbüchle, O., M. Santer, J. A. Steitz, and R. J. Mans.** 1978. Conservation of the primary structure at the 3' end of 18S RNA from eukaryotic cells. Cell **13**: 551-563.
26. **Hayward, G. S., and M. G. Smith.** 1972. The chromosome of bacteriophage T5. I. Analysis of the single-stranded DNA fragments by agarose gel electrophoresis. J. Mol. Biol. **63**:383-395.
27. **Hewick, R. M., M. Fried, and M. D. Waterfield.** 1975.

Nonhistone virion proteins of polyoma: characterization of the particle proteins by tryptic peptide analysis by use of ion-exchange columns. Virology **66**:408-419.

28. **Hewick, R. M., A. Mellor, A. E. Smith, and M. D. Waterfield.** 1980. Partial amino-terminal sequences of the polyoma nonhistone proteins VP1, VP2, and VP3 synthesized in vitro. J. Virol. **33**:631-636.
29. **Hewick, R. M., M. D. Waterfield, L. K. Miller, and M. Fried.** 1977. Correlation between genetic loci and structural differences in the capsid proteins of polyoma virus plaque morphology mutants. Cell **11**:331-338.
30. **Horowitz, M., S. Bratosin, and Y. Aloni.** 1978. Polyoma infected cells contain at least three spliced late RNAs. Nucleic Acids Res. **5**:4662-4674.
31. **Howley, P. M., M. A. Israel, M.-F. Law, and M. A. Martin.** 1979. A rapid method for detecting and mapping homology between heterogenous DNAs. Evaluation of polyoma virus genome. J. Biol. Chem. **254**:4876-4883.
32. **Hunter, T., and W. Gibson.** 1978. Characterization of the mRNA's for the polyoma virus capsid proteins VP1, VP2, and VP3. J. Virol. **28**:240-253.
33. **Kamen, R., J. Favaloro, and J. Parker.** 1980. Topography of the three late mRNA's of polyoma virus which encode the virion proteins. J. Virol. **33**:637-651.
34. **Kamen, R., J. Favaloro, J. Parker, R. Treisman, A. J. Flavell, A. Cowie, and S. Legon.** 1979. Structure of polyoma virus mRNAs synthesized in productively infected mouse cells. Differentiation **13**:45-46.
35. **Kamen, R., and H. Shure.** 1976. Topography of polyoma virus messenger RNA molecules. Cell **7**:361-371.
36. **Kimura, M.** 1968. Evolutionary rate at the molecular level. Nature (London) **217**:624-626.
37. **Klug, A.** 1965. Structure of viruses of the papilloma-polyoma type. II. Comments on other work. J. Mol. Biol. **11**:424-431.
38. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**:560-564.
39. **Miller, L. K., and M. Fried.** 1976. Construction of the genetic map of the polyoma genome. J. Virol. **18**:824-832.
40. **Miyata, T., and T. Yasunaga.** 1978. Evolution of overlapping genes. Nature (London) **272**:532-535.
41. **Pan, J., V. B. Reddy, B. Thimmappaya, and S. M. Weissman.** 1977. Nucleotide sequence of the gene for the major structural protein of SV40 virus. Nucleic Acids Res. **4**:2539-2547.
42. **Porter, A. G., J. Merregaert, J. Van Emmelo, and W. Fiers.** 1978. Sequence of 129 nucleotides at the 3'-terminus of encephalomyocarditis virus RNA. Eur. J. Biochem. **87**:551-561.
43. **Proudfoot, N. J., C. C. Cheng, and G. G. Brownlee.** 1974. Sequence analysis of eukaryotic mRNA. Prog. Nucleic Acids Res. Mol. Biol. **19**:123-134.
44. **Reddy, V. B., B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, P. K. Ghosh, M. L. Celma, and S. M. Weissman.** 1978. The genome of simian virus 40. Science **200**:494-502.
45. **Rogiers, R., A. van de Voorde, E. Soeda, and W. Fiers.** 1978. Nucleotide sequence of simian virus 40 Hind-K restriction fragment. Eur. J. Biochem. **85**:205-224.
46. **Russel, G. S., P. M. B. Walker, R. A. Elton, and J. H. Subak-Sharpe.** 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. **108**:1-23.
47. **Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith.** 1977. Nucleotide sequence of bacteriophage ϕX174 DNA. Nature (London) **265**:687-695.
48. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**:5463-5467.
49. **Seif, I., G. Khoury, and B. Dhar.** 1979. BKV splice sequences based on analysis of preferred donor and acceptor sites. Nucleic Acids Res. **6**:3387-3398.
50. **Shah, K. V., H. L. Ozer, H. N. Chazey, and T. J. Kelly, Jr.** 1977. Common structural antigen of papovavirus of the simian virus 40-polyoma subgroup. J. Virol. **21**:179-186.
51. **Siddell, S. G., and A. E. Smith.** 1978. Polyoma virus has three late mRNA's: one for each virion protein. J. Viroi. **27**:427-431.
52. **Smith, A. E., R. Kamen, W. F. Mangel, H. Shure, and T. Wheeler.** 1976. Location of the sequences coding for capsid proteins VP1 and VP2 on polyoma virus DNA. Cell **9**:481-487.

52a. **Soeda, E., J. R. Arrand, and B. E. Griffin.** 1979. Polyoma virus. The early region and its T-antigens. Nucleic Acids Res. **7**:839-857.

53. **Soeda, E., J. R. Arrand, N. Smolar, and B. E. Griffin.** 1979. Sequence from early region of polyoma virus DNA containing viral replication origin and encoding small, middle and (part of) large T antigens. Cell **17**:357-370.
54. **Soeda, E., G. Kimura, and K.-I. Miura.** 1978. Similarity of nucleotide sequence around the origin of DNA replication in mouse polyoma virus and simian virus 40. Proc. Natl. Acad. Sci. U.S.A. **75**:162-166.
55. **Subramanian, K. N.** 1979. Segments of simian virus 40 DNA spanning most of the leader sequence of the major late viral messenger RNA are dispensable. Proc. Natl. Acad. Sci. U.S.A. **76**:2556-2560.
56. **Walter, G., and W. Deppert.** 1975. Intramolecular disulfide bonds: an important structural feature of the polyoma virus capsid. Cold Spring Harbor Symp. Quant. Biol. **39**:255-257.