THE PHYSICAL AND GENETIC ORGANIZATION OF A VIRAL GENOME

Author: **Paul Berg**
Department of Biochemistry
Stanford University School of Medicine
Stanford, California

INTRODUCTION

What Harland Wood enjoys most, beside his family, deer hunting, and a raucous party, is good science: a bright idea, an interesting speculation, a well designed experiment, a clearly written paper, a sparkling talk. These are the things that turn Harland on. Clearly, the organizers have chosen the most appropriate means to honor Harland on his 70th anniversary: a stimulating program of lectures and the gathering of family, comrades in arms, and old friends.

I was especially pleased to be invited to speak on this occasion for it has given me an opportunity to tell Harland, in a way he appreciates, of my deep admiration and affection for him. I am one amongst many whose values in science and life were influenced by their encounter with Harland. In addition to his own example, Harland created an environment—a community of outstanding colleagues who shared his commitment and ideals—that encouraged students to share in the excitement of discovery. I, particularly, benefitted from close associations with Warwick Sakami and G. Robert Greenberg during my fumbling attempts to become a researcher. Their encouragement and advice, the friendship of Mert Utter, Victor Lorber, John Muntz, and Tom Singer and the high expectations they held for all of us made the introduction to serious science a memorable period in my life.

While I was a student in biochemistry at Western Reserve University (1948—1952), viruses were not much discussed in the classroom, seminars, or laboratory. I do recall, however, Jerry Hurwitz returning from a summer sojourn at Cold Spring Harbor and bubbling with enthusiasm about something he called phages (he pronounced it to rhyme with lodges). But most of the students (and I suspect the faculty as well) were oblivious to the growing importance of bacteriophages (I pronounce it to rhyme with cages) in the impending revolution in genetics. Now, 25 years later, it is clear that much of our sophistication about genetic chemistry rests upon advances made in the molecular biology of bacteriophages.[1] Both the DNA and RNA bacteriophages provided elegant models and reagents for analyzing the molecular details of gene expression, i.e., transcription, translation, and regulation; circular DNA phages permitted the dissection of *E. coli's* machinery for DNA replication (see Kornberg's and Hurwitz's papers in this volume); and phage models simplified the analysis of the molecular mechanisms of mutagenesis, recombination, radiation damage, and DNA repair.

The bacteriophages had such a pervasive influence because their considerably simpler genomes can be obtained in large quantities, and pure form. Consequently, they provide ideal substrates for enzymatic, structural, and nucleotide sequence studies, as well as reagents for monitoring their own genetic expression. Moreover, the synchronous introduction of exogenous viral chromosomes into cells facilitates the analysis of the regulatory mechanisms governing the virus' genomic replication and expression.

It is not surprising, therefore, that those undertaking studies on the molecular mechanisms of gene expression and regulation in higher organisms, particularly mammalian cells, turned to viruses as models. Over the years several viruses have proved to be extremely useful; perhaps, the most spectacular successes have been obtained with the DNA-containing adenoviruses and papova viruses (e.g., SV40 and polyoma) and several RNA-containing viruses (e.g., retro-, picorna-, and reoviruses). My own research efforts have used SV40 as the model and what follows summarizes the deductions that have been made about the genetic organization and expression of its chromosome.

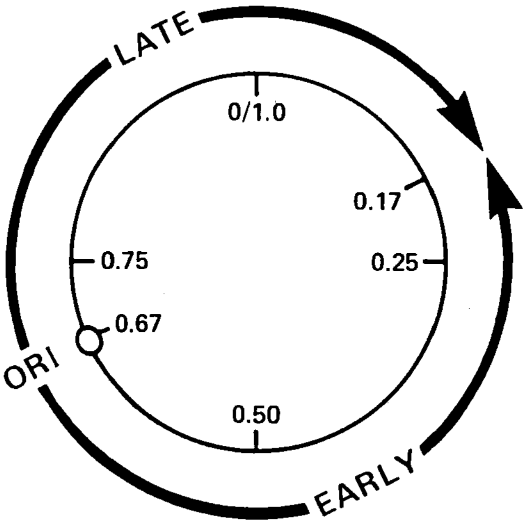
Why did we and many other laboratories adopt SV40 as a model for studying eukaryote gene expression? Initially, the virus attracted attention because it could induce tumors in rodents and transform normal cultured cells into tumor-like cells. Undoubtedly the challenge of discovering the molecular mechanism of viral oncogenesis was a lure, but the virus also has a vegetative life-style in which it kills its host and produces progeny virus. Thus, the SV40 genome contains a genetic program for producing virions; it also contains one or more genes that can alter the morphology and growth regulation of cells it infects. (There are several excellent reviews and a book that summarize the basic molecular and cellular aspects of SV40 biology. The references provided here to these sources[2-4] are intended to guide the interested reader further and to relieve me of the necessity of citing individual sources for documentation of general statements and widely accepted conclusions).

Structurally, SV40 is relatively simple: the nearly spherical particles consist of a protein capsid containing three viral-coded polypeptides; enclosed within the capsid is a chromatin-like structure composed of a covalently circular double-stranded DNA molecule of about 5200 base pairs (5.2 kilobase pairs [kb]) complexed with four host cell-derived histones, H2a, H2b, H3, H4. The 5.2 kb DNA molecule codes for three polypeptides that comprise the capsid and two proteins which are necessary for viral DNA replication and cellular transformation (T antigens). (The T antigen may be the precursor of another viral-coded polypeptide — the tumor-specific transplantation antigen [TSTA]). Our goal was to define each of the SV40 genetic elements, i.e., its structural and regulatory genes, and to map their physical location on the 5.2 kb DNA molecule.

Restriction endonucleases have played a crucial part in the efforts to define the genetic and functional organization of the SV40 genome. The restriction or cleavage sites provide coordinates for a molecular map of the DNA and permit one to locate, accurately, particular physical features or genetic loci. The single *Eco*RI endonuclease cleavage site serves as the reference point and is assigned map position 0/1.0 (Figure 1); all other positions in the DNA are given map coordinates in SV40 DNA fractional lengths, measured clockwise from 0/1.0 (Figure 1).

Following infection of permissive (primate) cells the SV40 genome is expressed in a temporally regulated order. Initially, RNA transcripts, complementary to one strand (the E strand) of about one half of the DNA, appear in cytoplasmic polysomes. The RNA transcripts, synthesized in the counterclockwise direction (see Figure 1), encode the structural information for the T antigens (and probably TSTA). Synthesis of T antigen triggers viral DNA replication beginning at map position 0.67 and proceeding bidirectionally terminating at about map coordinate 0.17. Concommitant with DNA replication, two, or perhaps three, additional viral RNA transcripts appear in the polysomes; these RNA molecules are transcribed in the clockwise direction being complementary to the other strand (the L strand) of the other half of the DNA (see Figure 1). The late mRNAs code for the synthesis of the capsid proteins VP1, VP2, and VP3. (VP1 is the major, and VP2 and VP3 the minor proteins in the SV40 capsid; VP2 and VP3 are related in that they share a rather extensive amino acid sequence which is distinct from the VP1 amino acid sequence). Synthesis of progeny DNA molecules and the capsid proteins results in death of the cell and production of mature virions.

When SV40 infects nonpermissive (nonprimate mammalian) cells the outcome is different. The same early events take place: the E strand transcripts and T antigens



VIRAL CODED PROTEINS

EARLY	
LARGE-T	(90-100 Kd)
Small-t	(15-20 Kd)
"TSTA"	(50-60 Kd?)
LATE	
VP-1	(42 Kd)
VP-2	(38 Kd)
VP-3	(25 Kd)

FIGURE 1. The functional map of the SV40 genome. The "early region" codes for the T antigens and, probably, the TSTA antigen; the "late region" codes for the capsid polypeptides VP1, VP2, and VP3. The numbers in parenthesis are the molecular weights of the proteins in kilodaltons estimated from their electrophoretic mobilities in denaturing gels. (From Berg, P., *Miami Winter Symp. XIII*, Scott, W. A. and Werner, R., Eds., Academic Press, New York, 1977. With permission.)

are made, but DNA replication, late strand transcription, and capsid protein synthesis do not occur. Generally, cell DNA replication and mitosis are induced, but most cells revert to a normal state and show no evidence of prior infection. A small proportion of the cells (less than 10%) become genetically transformed. These cells can continue to divide under culture conditions that restrict the multiplication of normal cells, they have an altered morphology, and can produce tumors when innoculated into appropriate animals. Significantly, the transformed cells contain all or part of the viral DNA, covalently integrated into the cell's chromosomal DNA.

When we began our work much of what I have just summarized was unknown. But Peter Tegtmeyer, at WRU, and later, Robert Martin at the National Institutes of Health (NIH) had isolated a variety of thermosensitive (ts) SV40 mutants; some were defective in both DNA replication and transformation; other mutants performed these functions normally but failed to make infectious virions. The Tegtmeyer and Martin mutants were sorted into different complementation and functional groups by genetic and physiologic tests, but mapping the mutational sites, and thereby the position of the genes coding for the various proteins, was not possible then. Accordingly, in 1972 we set out to isolate SV40 mutants with substantial alterations in their DNA structure (e.g., deletions, substitutions, or additions, etc.) and to map the position of these changes on the DNA molecule using physical and enzymatic means. We expected that deletions in structural genes would cause easily demonstrable changes in their polypeptide products, thereby permitting us to identify and map the structural genes coding for particular polypeptides.

During the past five years a variety of procedures for the construction, in vitro, of SV40 mutants with deletions of as little as three to five base pairs to as many as several hundred to a thousand base pairs have been developed in our laboratory. Similarly, techniques for the sizing and mapping of these deletions were worked out, permitting us to define the exact position of each alteration on the DNA molecule. Some of the mutants are viable, that is, they are still able to grow in cultured monkey cells without need for any helper virus; presumably, no vital function has been inactivated. Most mutants, however, are defective and nonviable; these can be grown in the presence of a helper virus which provides the function that has been lost in the mutant. Since the essential features of these procedures have been summarized recently,[5,6] I shall not discuss them here. Instead, I intend to focus on what deductions can be drawn from our studies, on the organization of SV40 genes.

Figure 2 shows a map of SV40 which indicates the DNA regions within which deletions have been obtained. (The map coordinates are given in SV40 fractional length, measured clockwise from map coordinate 0/1.0, the *Eco*RI, endonuclease cleavage site; 0.01 map unit equals 52 base pairs). The first point to note is that no deletion, amongst the many that have been examined, removes map coordinate 0.67, the origin of DNA: replication (ORI). That is not surprising since the isolation of a mutant genome requires that it be able to replicate. However, since there are mutants whose deletions extend to within 50 base pairs of coordinate 0.67, it seems clear that the nucleotide information needed to specify the SV40 origin of DNA replication is contained within the 100 or so base pairs between map coordinates 0.66 and 0.68. Further studies and alternative approaches are needed to define more precisely what special structural features serve as this minichromosome's DNA replication initiation site.

Considering first the "late region" (the DNA segment clockwise, between map coordinates 0.67 and 0.17), there is a region from ORI to about map position 0.75 which does not code for any virion capsid protein, or for that matter any known viral protein. Some 20 mutants with deletions (and some with insertions) ranging in size from 20 to 200 base pairs within the DNA segment defined by map coordinates 0.68 to 0.75 have been examined; each grows without a helper virus and produces apparently normal capsid proteins. Mutants carrying deletions within this region grow more slowly than wild type virus but the reason for this is unclear.

Let me skip, for the moment, the region 0.76 to 0.79, and consider a set of mutants with deletions that occur between map positions 0.80 and 0.83. These mutants do not make normal viral capsids because they produce no, or only a defective VP2 polypeptide; however these mutants synthesize normal amounts and sized VP1 and VP3 polypeptides. Mutants with deletions that occur distal to 0.83 (several that lie between 0.80 and 0.90 are not shown on the diagram) affect the structure of both VP2 and VP3 polypeptides. Two are particularly informative: one mutant that has a deletion of about 50 bp at map position 0.93 produces VP2 and VP3 polypeptides that are 6 kilodaltons (kd) shorter than the wild type polypeptide; another mutant, with a small deletion at 0.94, also produces smaller (about 4 kd) VP2 and VP3 polypeptides. Both mutants produce normal sized VP1. From these findings we infer that the structural

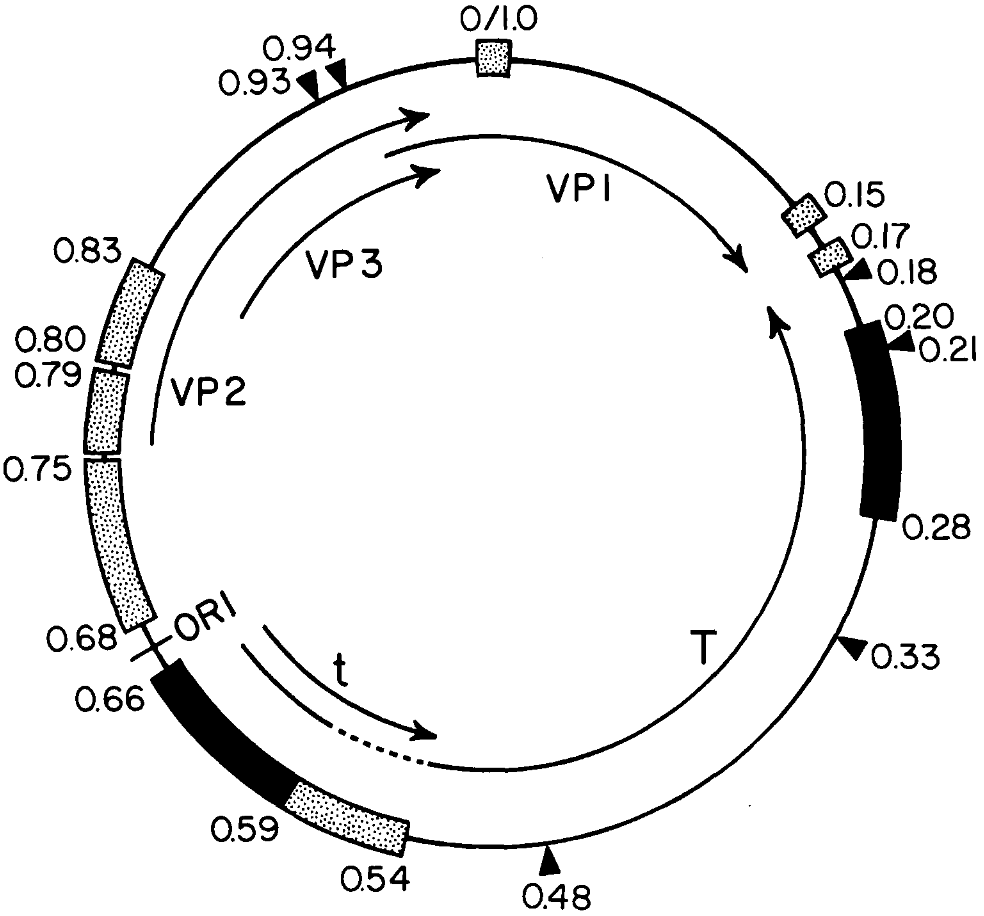


FIGURE 2. Map location of SV40 deletion mutants on the SV40 chromosome. Coordinates are in SV40 DNA fractional length, measured clockwise from the *Eco*RI endonuclease-cleavage site at 0/1.0. The arrows indicate the positions of the coding sequences for the late proteins VP1, VP2, and VP3, and the two documented early proteins, large T and small t. The stippled areas indicate segments within which more than one mutant is known, the solid bars represent single extended deletions and the small triangles represent single deletions at that particular map location. (From Cole, C. N., Landers, T., Goff, S. P., Manteuil-Brutlag, S., and Berg, P., *J. Virol.*, 24, 277, 1977. With permission.)

genes for VP2 and VP3 are contained within the DNA segment from map position 0.75 to 1.0, that the region coding for VP2 begins between map position 0.76 and 0.79, and that the region coding for the common amino acid sequence of VP2 and VP3 is distal to map position 0.83 and extends beyond coordinate 0.94.

Deletions at the *Eco*RI endonuclease-cleavage site (map position 0/1.0) and at map coordinate 0.14 affect the structure and function of VP1, whereas deletions at 0.16 to 0.18 have no effect on the structure or function of VP1. None of these deletions affect the structure of VP2 and VP3 polypeptides. Thus, the gene coding for VP1 begins beyond map position 0.94 but before 1.0 and extends beyond map position 0.14 and terminates before 0.16.

A more refined definition of the coding sequences for the capsid's three polypeptide chains has been obtained from an examination of the nucleotide sequence of SV40 DNA.[7,8] The most probable AUG codons for initiating the synthesis of VP1, VP2, and VP3 are located at map positions 0.945, 0.775, and 0.835, respectively. The termination codon for VP1 occurs at about map position 0.15, accounting for its estimated molecular weight of 42 kd. The position of the first in-phase terminator for VP2 and VP3 was unexpected. Instead of occurring before the start codon for VP1, the terminator codon for VP2 and VP3 is beyond the VP1 initiator AUG; in fact, the translation of VP2 and VP3 continues 121 nucleotides past the VP1 initiator codon, albeit in another phase. This explains why the deletions at map positions 0.93 and 0.94, which were thought to be so close to the start of the VP1 gene, cause VP2 and VP3 to be as much as 6 and 4 kd, respecitvely, shorter than the wild type polypeptides. Using these coordinates for the VP2 and VP3 coding sequences makes their mol wt 38 to 40 kd and 28 to 30 kd, respectively.

A particularly interesting group of mutants are those with deletions that occur within, or extend into, the DNA segment bounded by map coordinates 0.75 to 0.79. These mutants are unable to produce VP1, VP2, and VP3 polypeptides. This was unexpected because, although such deletions invade the structural gene for VP2, they clearly leave the coding sequences for VP3 and VP1 unaltered. Evidently, a nucleotide sequence between 0.75 and 0.79 is essential for proper expression of distal genetic information. In view of the fact that the mRNAs coding for VP1, VP2, and VP3 contain, in addition to their protein coding sequences, a 5′-"leader" sequence of about 150 to 200 nucleotides complementary to DNA sequences from map position 0.72 to 0.76,[9-11] it is quite possible that the 0.75 to 0.79 region is involved in generating the properly "spliced" late mRNAs.

The analysis of mutants with deletions in the early region proved to be equally revealing. At the time we began our studies, the generally accepted view was that the SV40 early region coded for a single protein (T antigen) of mol wt 90 to 100 kd; this protein was known to be essential for both viral DNA replication and oncogenic transformation.[2-4] To code for a polypeptide of that molecular weight, virtually the entire early region would be needed. Yet a prominent class of our early mutants, those with deletions between map positions 0.59 to 0.54, grows nearly as well as wild type virus in the absence of a helper; moreover, cells infected with such mutants produce a normal-sized T antigen. Since mutants with deletions between map coordinates 0.54 and 0.20 were generally defective and produced shortened T antigens, we surmised that the coding sequence for T antigen was beyond map position 0.54; and, since deletions at map position 0.16 to 0.17 did not alter or inactivate T antigen, the other boundary of the structural gene was placed proximal to map position 0.17. By our estimates T antigen could not be larger than 65 to 70 kd! This view was supported by nucleotide sequence data which revealed the existence of terminator codons in all three possible translation phases at about map position 0.55.[7,8]

But how could we explain the finding that another mutant, lacking the DNA segment between map position 0.66 and 0.59 was not viable and could not produce a functional T antigen? The key to the puzzle came from the discovery[12,13] that the SV40 early region codes for two antigenically related polypeptides—large T and small t, whose molecular weights are 90 to 100 kd and 15 to 100 kd, respectively. Particularly significant was the finding[12,13] that small t and large T share a common set of tryptic peptides, suggesting that they have a common amino acid sequence and probably a common coding sequence. To determine which part of the early region codes for small t, we reexamined our collection of deletion mutants for their ability to make small t and large T.[14] To our surprise, only the mutants with deletions in the 0.59 to 0.54 region were defective in their ability to make small t. This made the puzzle even more intriguing. For how could deletions alter small t structure and leave large T unaffected if the two are coded by a common nucleotide sequence? To account for this we have suggested that large T and small t are coded by two different early mRNAs; the mRNA coding for small t is homologous to the entire early region (from map position 0.67

counterclockwise to 0.17) and the mRNA coding for large T spans the same region but lacks the sequences present in the DNA segment between map positions 0.59 and 0.54.[14]

Translation of the larger mRNA, the one containing the 0.59 to 0.54 region, begins with an AUG at map position 0.65 and terminates at a UAA at 0.545 to produce small t. Translation of the shorter mRNA, the one lacking the 0.59 to 0.54 region, begins at the same AUG and proceeds unimpeded to the first in phase terminator codon at map position 0.175 yielding large T. The existence of two early mRNAs, one of which lacks the nucleotide sequences between map positions 0.59 and 0.54, has been confirmed experimentally,[15] although the mechanism for producing mRNAs lacking internal nucleotides remains a mystery. Further support for the proposed model comes from the finding that the NH_2 terminal thirty or so amino acids of large T and small t are identical.[16] Further work is needed to determine at which point in the two proteins the amino acid sequences diverge and how spliced mRNA sequences are created.

The identification of small t as a new SV40-coded protein raises questions as to its function. Initially, when these mutants were isolated we considered that some deletions might inactivate the virus oncogenic potential without affecting its ability to grow. But our test of the mutant's ability to transform nonpermissive cells showed that they were normal. Recently, in collaboration with Giampiero di Mayorca's laboratory, this question was reinvestigated.[17] Using a more stringent test of oncogenic transformation, i.e., a cell's ability to produce a colony in soft agar,[18] the mutants defective in small t formation were only 1 to 2% as effective as wild type virus.[6] Mutants that fail to make normal large T antigen are also defective for cellular transformation. But large T and small t must perform separate functions in transformation as the two types of mutants complement each other for transformation.[17]

The striking similarity in the physical and functional organization of the SV40 and polyoma genomes has long been apparent;[2,3] this similarity now extends to the existence and function of the beginning of the early region. Benjamin's laboratory[19] has noted that a class of polyoma mutants, which can grow in some hosts but not others, is unable to transform nonpermissive cells; several of the mutants of this class have deletions at the beginning of the early region.[19] The recent finding[20] that one of the mutants in this class, NG-18, fails to produce a normal small t makes the analogy even closer. The intriguing questions of how small t and large T function in oncogenesis remain to be elucidated.

REFERENCES

1. **Cairns, J., Stent, G. S., and Watson, J. D., Eds.,** *Phage and origins of molecular biology,* Cold Spring Harbor Laboratory, New York, 1966.
2. **Tooze, J., Ed.,** *The molecular biology of tumour viruses,* Cold Spring Harbor Laboratory, New York, 1973.
3. **Kelly, T. J., Jr. and Nathans, D.,** The genome of simian virus 40, *Adv. Virus Res.,* 21, 85, 1977.
4. **Fried, M. and Griffin, B. E.,** Organization of the genomes of polyoma virus and SV40, *Adv. Cancer Res.,* 24, 67, 1977.
5. **Berg, P.,** The Eighth Feodor Lynen Lecture: biochemical pastimes and future times, in *Miami Winter Symp. XIII,* Scott, W. A. and Werner, R., Eds., Academic Press, New York, 1977.
6. **Cole, C. N., Landers, T., Goff, S. P., Manteuil-Brutlag, S., and Berg, P.,** Physical and genetic characterization of deletion mutants of simian virus 40 constructed *in vitro, J. Virol.* 24, 277, 1977.
7. **Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, H., Van Houverswyn, J., Van Herreweghe, G., Volckaert, G., and Ysebaert, M.,** The total nucleotide sequence of SV40 DNA, *Nature (London),* 273, 113, 1978.

8. **Reddy, V. B., Thimmappaya, B., Dhar, R., Subraminian, K. H., Zain, B. S., Celma, M. L., Pan, J., and Weissman, S. M.**, The genome of simian virus 40, *Science*, 200, 494, 1978.
9. **Aloni, Y., Dhar, R., Laub, O., Horowitz, M., and Khoury, G.**, Novel mechanism for RNA maturation: the leader sequences of simian virus 40 mRNA are not transcribed adjacent to the coding sequences, *Proc. Natl. Acad. Sci., U.S.A.*, 74, 3686, 1977.
10. **Hsu, M. -T. and Ford, J.**, Sequence arrangement of the 5′-ends of simian virus 40 16S and 19S mRNAs, *Proc. Natl. Acad. Sci., U.S.A.*, 74, 4982, 1977.
11. **Lavi, S. and Groner, Y.**, 5′-Terminal sequences and coding region of late simian virus 40 mRNAs are derived from noncontiguous segments of the viral genome, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 5323, 1977.
12. **Prives, C., Gilboa, E., Revel, M., and Winocour, E.**, Cell-free translation of simian virus 40 early messenger RNA coding for viral T antigen, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 457, 1977.
13. **Paucha, E., Harvey, R., Smith, R., and Smith, A. E.**, The cell-free synthesis of SV40 T-antigens, *INSERM Collo.*, 69, 189, 1977.
14. **Crawford, L. V., Cole, C. N., Smith, A. E., Paucha, E., Tegtmeyer, P., Rundell, K., and Berg, P.**, Organization and expression of early genes of simian virus 40, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 117, 1978.
15. **Berk, A. J. and Sharp, P. A.**, Spliced early messenger RNAs of simian virus 40, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 1274, 1978.
16. **Paucha, E., Mellor, A., Harvey, R., and Smith, A. E.**, SV40 large and small T-antigens have identical amino termini mapping at 0.65 map units, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
17. **Bouck, N., Beales, N., Shenk, T., Berg, P., and di Mayorca, G.**, New region of the simian virus 40 genome required for efficient viral transformation, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 2473, 1978.
18. **Shin, S. I., Freedman, V. H., Risser, R., and Pollock, R.**, Tumorgenicity of virus-transformed cells in nude mice is correlated specifically with anchorage independent growth *in vitro*, *Proc. Natl. Acad. Sci. U.S.A.*, 72, 4435, 1975.
19. **Benjamin, T. L.**, HR-T mutants of polyoma virus, in *Miami Winter Symp., XIV*, Schulte, J. and Buada, Z., Eds., Academic Press, 1977.
20. **Ito, Y., Brockelhurst, J. R., Spurr, N., Griffiths, M., Hurst, J., and Fried, M.**, INSERM Colloq. (Paris), 69, 69, 1977.