*Proc. Natl. Acad. Sci. USA*
Vol. 77, No. 6, pp. 3302–3306, June 1980
Biochemistry

# Nucleotide sequence of Moloney leukemia virus: 3′ End reveals details of replication, analogy to bacterial transposons, and an unexpected gene

(reverse transcription/inverted repeats/p15E/*env* gene/mink cell focus-forming virus)

J. GREGOR SUTCLIFFE*, THOMAS M. SHINNICK*, INDER M. VERMA†, AND RICHARD A. LERNER*

*Department of Cellular and Developmental Immunology, Scripps Clinic and Research Foundation, La Jolla, California 92037; and †The Salk Institute, San Diego, California 92112

*Communicated by Frank J. Dixon, March 20, 1980*

ABSTRACT **We have determined the sequence of a cloned DNA fragment 1108 base pairs long which corresponds to the 3′ end of the Moloney murine leukemia provirus. The clone was obtained as the primary product of reverse transcription and begins with the Moloney "strong stop" sequence, then extends towards the 5′ end of the provirus. Our sequence: (*i*) proves that reverse transcriptase switches templates during minus strand synthesis; (*ii*) defines the limits of the 515-base-pair repeats which occupy both ends of the integrated provirus; (*iii*) shows that the structure of the proviral repeats has strong analogy to bacterial insertion sequences, indicating that the Moloney provirus is a transposon; (*iv*) identifies the putative promotor for genomic transcription within these repeats; (*v*) shows that the presumed origin of second strand synthesis, which lies just outside the 3′ repeat, has tertiary structure analogous to single-stranded bacteriophage origins of replication; (*vi*) solves the amino acid sequence of most of p15E, the carboxy-terminal product of the *env* gene; (*vii*) allows detailed mapping of the mink cell focus-forming virus substitution locus in a central location within the gp70 region of the *env* gene; and (*viii*) identifies a long open translation frame to the right of the *env* gene (*R* gene) which could be involved in leukemogenesis.**

Moloney murine leukemia virus (MLV) (1) has long been the focus for biochemical and structural studies as a prototype for the mammalian RNA tumor viruses. Central to understanding the biology of retroviruses is the unambiguous assignment of genes to the RNA genome. In addition, because these viruses represent fragments of "selfish" nucleic acid (2) which can exist either free in a virion or associated with a cellular genome, the ends of the replicating molecule and its mode of replication are of particular interest.

After infection, the virus-coded reverse transcriptase enzyme copies the Moloney virus single-stranded RNA genome into double-stranded DNA (3). Studies of reverse transcription *in vitro* indicate that DNA synthesis is initiated by the covalent elongation of the priming tRNA molecule which is bound near the 5′ end of the RNA genome (4). Synthesis proceeds in a 5′-to-3′ fashion, polymerizing deoxyribonucleotides complementary to the RNA genome such that the 5′ end of the template is soon reached. The nascent single-stranded DNA molecule is then thought to migrate to the 3′ end of the RNA template, where it may pair by virtue of complementary nucleotides with the string of approximately 60 bases immediately 5′ to the poly(A) tail on that end of the genome (5). The nascent chain is then elongated, presumably continuously, to the 5′ end of the template, producing a complete minus strand DNA copy of the virus genome. The mechanism for second-strand DNA synthesis and for producing a repeated end structure, thus forming the double-stranded DNA molecule, has not yet been so clearly elucidated. The double-stranded DNA molecule with repeated ends, then, is the structure that may integrate into the host chromosome, where it is inherited in a Mendelian fashion and may act as a substrate for transcription and thus formation of new virus particles.

We are studying the structure of proviral DNA cloned in bacterial vectors. Our initial attention has been drawn to the 3′ end, where much of the molecular interest is. We choose to study the nucleotide sequences of cloned DNA fragments rather than that of whole viral RNA preparations because these RNA viruses have a low ratio of infectious virus to particle and the RNA is quite heterogeneous. Ultimately, the cloned DNA segments may be shown by transfection studies to harbor full biological activity, a feature not possible with RNA populations.

We have investigated the molecular architecture of the product of the initial events in reverse transcription and have completely defined at the DNA sequence level the 5′ end of the minus strand DNA. The sequence extends into the carboxy-terminal end of the most 3′-proximal gene of the viral RNA. Our study proves the molecular jump (4, 6–9), identifies important transcription and replication signals, defines the insertion sequence that indicates that the provirus is a transposable element, and maps the gene for p15E on the virus genome, as well as provides most of the p15E amino acid sequence. This mapping shows that p15E is a virus-coded product and allows the location of the *env* gene to be calculated. In addition, we report here the presence of a long open reading frame to the right of the *env* gene and discuss its possible relevance to leukemogenic transformation.

## MATERIALS AND METHODS

The source of this cloned, plasmid DNA has been described (10). Briefly, cDNA was made *in vitro* by detergent-disrupted, purified Moloney murine leukemia virus (11); after treatment by the single-strand-specific nuclease S1, it was size-enriched on a sucrose gradient. dC residues were added to the ends of this double-stranded DNA with terminal transferase and cloned at the single *Pst* I site of pBR322, which has dG residues at its ends (12). The restriction profile of tetracycline-resistant, ampicillin-sensitive, hybridization-positive recombinant plasmids was compared to those of MLV DNA synthesized *in vitro* (13) and of pBR322 (14). A clone, pMLV-201, which was thought to contain the 5′-3′ jump, was chosen for sequence analysis. The sequence of restriction fragments was determined by the partial chemical degradation method (15, 16).

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: MLV, murine leukemia virus; $IR_R$ and $IR_L$, inverted repeat, right and left, respectively; MCF virus, mink cell focus-forming virus.

## RESULTS AND DISCUSSION

**Reverse Transcriptase Switches Templates.** Fig. 1 presents the 1195-base-pair sequence of this Moloney leukemia virus cDNA clone. Several features of the sequence lead us to believe that the recombinant plasmid is a faithful product of our described protocols. As expected, the insert is flanked by regenerated *Pst* I sites and has homopolymeric "tails" of 31 and 46 base pairs on the opposite ends of the insert. The recombinant plasmid hybridizes with MLV [$^{32}$P]cDNA and in fact begins at one end with that DNA sequence that corresponds to the so-called Moloney "strong stop" fragment, indicating that the cloned DNA was originally primed by tRNA. The restriction map of the insert is in general agreement with the published maps for Moloney cDNA molecules (13, 17). The length of the insert minus the tails added by the cloning procedure is 1108 nucleotide pairs. The bases are numbered starting with the first base after the tail.

The pMLV-201 insert sequence begins with a run of 143 base pairs that corresponds to the Moloney MLV analogue of the strong stop fragment first described in sequence detail for various avian retroviruses (6, 18, 19). The strong stop molecular species accumulates at a kinetic bottleneck when reverse transcripts are generated *in vitro* at low nucleotide triphosphate concentrations. This portion of the cDNA molecule, diagrammed in Fig. 2, is that which begins adjacent to the priming tRNA molecule and extends to the 5′ end of the RNA genome. Our sequence matches the Moloney strong stop sequence (20). The 143rd base, C, is complementary to the methyl-G identified as the 5′ ultimate ribonucleotide in continuous phosphodiester linkage with the body of the RNA genome (21). Curiously, this clone is missing the two A·T pairs that have previously been observed as the first two following the tRNA primer.

The cDNA molecule then jumps from 5′ material to 3′ material, such that nucleotides 144–1108 were copied by using the 3′ end of the genome as template. Actually, the 60 bases pre-

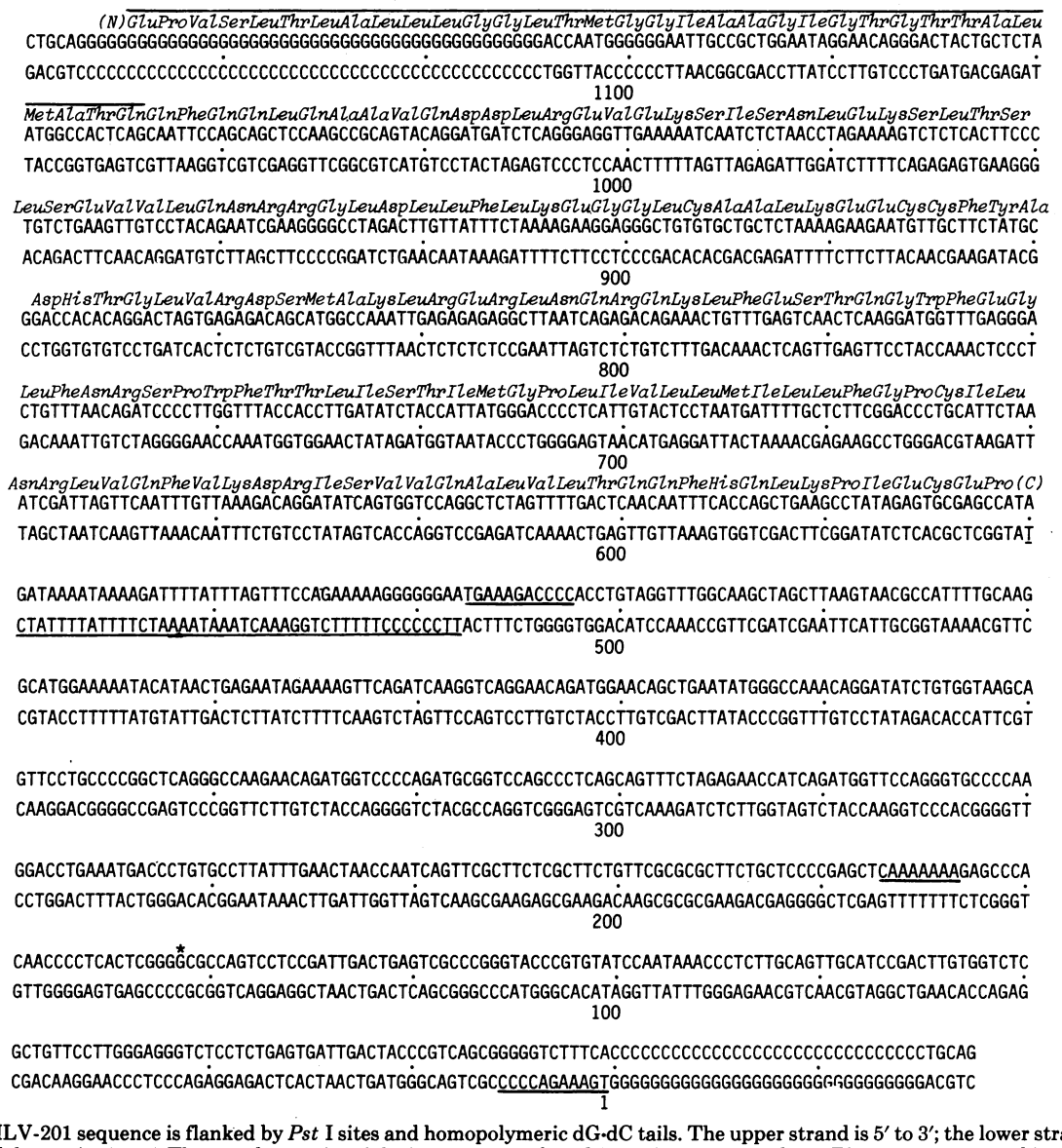


FIG. 1. pMLV-201 sequence is flanked by *Pst* I sites and homopolymeric dG·dC tails. The upper strand is 5′ to 3′; the lower strand is complementary and, hence, is 3′ to 5′. The 1108 base pairs of the insert are numbered from the first insert base. The insert is oriented in the pBR322 $\beta$-lactamase gene such that base 1 of the insert is nearer the $\beta$-lactamase amino terminus (16) than is base 1108. $IR_R$ and $IR_L$ (the 11-base-pair inverted terminal repeats), which flank the proviral repeats, are underlined (positions 1–11 and 505–515), as are the putative promotor (positions 167–174) and the origin of second-strand synthesis (positions 516–560). The amino acid sequence translated from the open reading frame is in italics above the nucleotide sequence; that from Oroszlan's protein sequence is overlined. The poly(dG) tail is not translated. The genomic 5′ base is denoted by an asterisk.

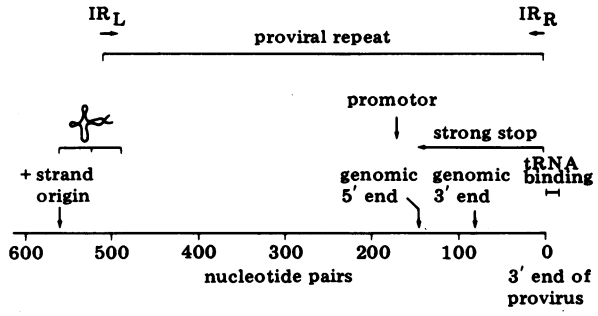


FIG. 2. Long terminal repeat of Moloney MLV is flanked by inverted terminal repeats ($IR_L$ and $IR_R$). The 5′ long terminal repeat contains an active promotor and the 5′ RNA genomic end. These are mirrored in the 3′ long terminal repeat which contains the 3′ RNA genomic end. The origin of second-strand synthesis and the stem and loop structure of Fig. 3 are represented.

ceding base 144 are represented at the 3′ as well as the 5′ end of the genome (8), but they are present only once on this cDNA molecule. We know that this region is the 3′ end of the viral genome because of the location of specific restriction cleavage sites (namely, *Xba* I and *Pvu* II) and the absence of sites for other cleavages. More compellingly, further down the molecule we encounter the *env* gene which is thought to be located near the 3′ end. Our sequence provides the necessary proof that the nascent strong stop intermediate actually switches ends to allow reverse transcriptase to copy the entire genome.

**MLV Is a Transposon.** Bases 1–515 represent one of the proviral repeats, the other copy of which appears at the 5′ end of the fully reverse transcribed provirus (22). In the accompanying paper, van Beveren *et al.* (20) demonstrate that the sequence at the 5′ end of the integrated Moloney MLV diverges completely from our 3′ sequence at base 516 (our numbers) but matches it in the bases preceding and including 515. In their case, bases 516 and on represent cellular DNA; in our case these sequences are in the body of the virus. This comparison defines the size of the repeat at 515 base pairs.

At either end of the proviral repeat we find a completely homologous sequence of 11 base pairs in inverted orientation

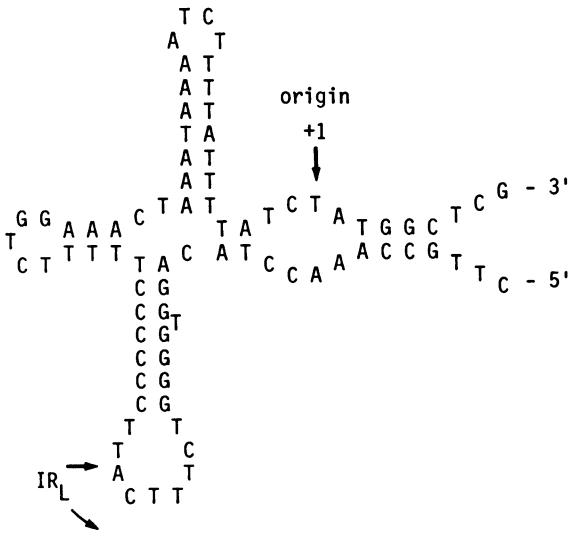


FIG. 3. Origin of second-strand synthesis can be drawn as a hairpin structure. These bases are from the minus strand, positions 487–568. The inverted repeat $IR_L$ is indicated. The position of initiation (+1) was identified by analogy to bacteriophage origins (26), but stable product probably begins at $IR_L$. Synthesis proceeds through the loops, complementary to these bases.

(underlined in Fig. 1 and diagrammed in Fig. 2). We have called them $IR_R$ and $IR_L$ for inverted repeat, right and left. The same repeating units are seen in the 5′ long terminal repeat (20). Such repeats are reminiscent of the structures found in genetic elements called insertion sequences (23). Because the body of the virus is flanked by two insertion elements (the provirus repeats) when the provirus is synthesized, it becomes a transposon, perhaps explaining, by analogy to prokaryotic transposons, how it integrates.

**Putative Promotor for Genomic Transcription.** Because the genome is thought to be a primary transcription product that is translated as an unspliced mRNA, it ought to have a promotor. The region that corresponds to the 5′ end of the RNA is represented at both ends of the provirus; thus the transcription initiation sequence at position 143 (on the lower strand) of this 3′ clone is embedded in the same sequence that the initiator of genomic transcription occupies. By analogy with other eukaryotic transcription units, the promotor should lie upstream 20–25 positions from position 143 (24). Underlined in Fig. 1 we see the sequence C-A-A-A-A-A-A-A at this location; an unimpressive five out of eight match with the customary T-A-T-A-A-A-T-A "Hogness box" (24). Our sequence differs by a single base from that of van Beveren at the third A. The difference does not make a better promotor fit. We think that this and two other single-base differences in the two sequences of the proviral repeats arise from errors made by the reverse transcriptase, which has an error rate of about 1/500.

Alternative mechanisms of RNA initiation can be considered, such as downstream promotion as is observed for the 5S genes of *Xenopus* (25). We do not find a sequence downstream from position 143 (positions 130–40, lower strand, were scanned) that shows any striking homology with the 5S internal promotor, and other possibilities seem less likely than using the C-A-A-A-A-A-A-A sequence, even though it poorly matches the Hogness box.

**Origin of Second-Strand Synthesis.** The sequence just into the body of the virus from the repeat junction, roughly positions 516–560 (underlined in Fig. 1), is of peculiar composition: on the upper strand, there is a sequence containing 15 pyrimidines in a row, then a 15-nucleotide sequence containing 12 purines, followed by an 11-nucleotide sequence containing 10 pyrimidines. This striking sequence occurs at the repeat junction, where the origin of second-strand replication is thought to be localized (22). A stem-and-loop structure (hairpin) may be drawn with these nucleotides (Fig. 3) similar to that for the single-stranded bacteriophage origins of replication (26). At the stage of reverse transcription at which this origin is used, the replicating DNA is single stranded (RNase H having removed the complementary RNA) and hence analogous to the small single-stranded phages. Analogy would place the second-strand start at about position 560. However, we notice that synthesis beginning at position 515 would generate the expected long terminal repeat. We suggest that an RNA primer similar to that used in G4 replication is synthesized *de novo* between positions 560 and 515 and that stable product begins at position 515. This model is consistent with the observed sensitivity of this step to actinomycin D.

**Gene for p15E.** When we examined the three possible triplet reading frames on both the plus and minus viral stands, we found nonsense termination codons frequently interrupting all but one translation frame. There are no open reading frames on the minus strand of DNA. The positive "sense" strand is open in one frame from the end of the clone (position 1108) to position 562. The protein sequence predicted by the DNA is presented in Fig. 1. Residues 2–20 match those determined by Oroszlan for positions 15–34 of p15E. We conclude that these

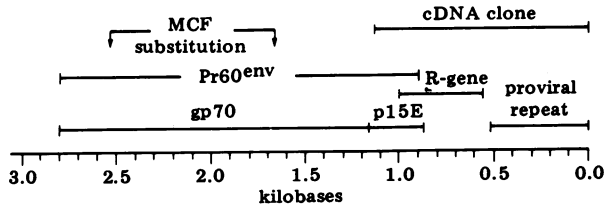


Fig. 4. Products of the *env* gene, gp70 and p15E, and their unglycosylated precursor, $Pr60^{env}$, are aligned on the proviral map, measured from the 3′ end, along with other markers. The position of MCF virus substitution in gp70 is indicated, as is the extent of our sequenced cloned DNA fragment. The proviral repeat is shown in more detail in Fig. 2.

amino acids must contain the carboxy-terminal portion of p15E and that the rest of the amino-terminal coding sequence lies just beyond the reach of our clone in the next 42 base pairs. The exact location of the p1SE carboxy terminus remains unidentified, but we estimate it to be between positions 850 and 900. This assignment of p15E maps its location on the genome and demonstrates that the protein is virus coded. Interestingly, the dominant region of heteroduplex nonhomology between Moloney and AKV viruses is located 900–1200 nucleotides from the 3′ ends of these viruses (27) and, hence, maps in the p15E coding region.

**This Frame Could Encode an Unidentified Protein.** The remainder of this open frame, from approximately position 880 to position 562, encodes a previously unrecognized product. We provisionally name this the R-protein to denote that it is encoded by the rightmost gene of the virus. If the *R* gene indeed encodes a stable product, then all of the DNA between the p15E gene and the proviral repeat is consumed by the functions of coding this protein and originating second-strand synthesis. p15E is one of the products of the *env* gene; thus the *R* gene is the last possible shell at this end of the genome under which a leukemogenic function could hide.

The *R*-gene amino acid sequence contains an extremely hydrophobic region, suggesting that the structure has membrane function, which could explain selective proliferation of certain differentiated cell types. If this implication is correct, one might expect viruses such as Friend and mouse mammary tumor, which transform other differentiated cell types, to encode in the *R*-gene region a protein of different specificity. Alternatively, the *R*-gene product could be involved in viral replication or transposition.

**Mapping the *env* Gene.** The product of the Moloney *env* gene is a polyprotein of 60,000–70,000 daltons carrying both gp70 (45–50 kilodaltons) and p15E (10–15 kilodaltons) peptides (28, 29). Pactymycin experiments determined gp70 and p15E at the amino and carboxy termini, respectively (30). Our sequence confirms that assignment and places the amino terminus of p15E, and hence the *env* precursor, at position 1150. This gp70 protein requires a gene of 1400–1650 nucleotides; hence the gp70 coding sequence spans from about position 1150 (COOH) to 2550–2800 ($NH_2$).

Leukemogenic mink cell focus-forming (MCF) viruses (31) differ from their more benign relatives by genetic substitutions which have been mapped by heteroduplex analyses (27, 32). The endpoints of these substitutions occur between 1.6–1.9 and 2.5–2.7 kilobases from the 3′ end, clearly disturbing the *env* gene, in agreement with the tryptic mapping results of Elder *et al.* (33). Fig. 4 shows these assignments. Recent studies by H. L. Niman and J. H. Elder (personal communication) using monospecific antibodies indicate that the MCF substitutions do not alter the amino-terminal approximately 20,000 daltons of gp70. Taken together, these data indicate that the MCF substitution is centrally located in the gp70 molecule. gp70 coding sequences must occupy most, if not all, of the 2800 nucleotides in the 3′-derived portion of the *env* 21S mRNA (32). This calculation shows that it is possible, although not necessary, that some, though not many, gp70 codons come from the 5′-derived spliced exon (300–600 nucleotides) (32). Ultimate resolution of this point relies on further sequence analysis.

The central location of the MCF *env* substitution in gp70 makes it likely that the recombinants are formed by a double crossover event. One mechanism for such an event could be a legitimate recombination between generally homologous regions of genes. Alternatively, we imagine that the exchange of central *env* regions could be mediated by a cassette mechanism similar to that shown to operate for yeast mating type (34) and immunoglobulin (35, 36) polymorphisms. If so, then MCF substitution sites will be found to be specific.

To invoke the *R* gene discussed above in murine leukemia, we must deal with the MCF viruses because changes 1000 nucleotides upstream from the *R* gene are highly correlated with viral leukemogenic potential (33). We suggest that there is coordination between gp70 and *R* and can imagine three classes of models that pertain. There could be a protein–protein interaction between gp70 and the *R* product. A second model requires that certain substitutions in the MCF region of *env* potentiate expression of the *R*-gene mRNA. Finally, the "passport" model has altered gp70 molecules expanding the host range of otherwise relatively benign viruses to include cell types that may respond to the *R*-gene product.

We thank J. Okamoto and R. Pesin for technical assistance and Drs. J. E. Elder, H. Niman, and S. Oroszlan for communicating results prior to publication. We gratefully acknowledge the Jane Coffin Childs Memorial Fund for Medical Research (J.G.S.) and the Helen Hay Whitney Foundation (T.M.S.) for fellowship support. This work was supported by National Institutes of Health Grant R01 CA 25325. This is publication no. 164 from the Department of Cellular and Developmental Immunology and no. 2079 from the Research Institute of Scripps Clinic.

1. Moloney, J. B. (1960) *J. Natl. Cancer Inst.* **24**, 993–952.
2. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284**, 604–607.
3. Verma, I. M. (1977) *Biochim. Biophys. Acta* **473**, 1–37.
4. Taylor, J. M. & Illmensee, R. (1975) *J. Virol.* **16**, 553–558.
5. Coffin, J. M. (1979) *J. Gen. Virol.* **42**, 1–26.
6. Haseltine, W. A., Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 989–993.
7. Schwartz, D. E., Zamecnik, P. C. & Weith, H. L. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 994–998.
8. Coffin, J. M., Hageman, T. C., Maxam, A. M. & Haseltine, W. A. (1978) *Cell* **13**, 761–773.
9. Peters, G., Harada, F., Dahlberg, J., Panet, A., Haseltine, W. & Baltimore, D. (1977) *J. Virol.* **21**, 1031–1042.
10. Sutcliffe, J. G., Shinnick, T. M., Lerner, R. A., Johnson, P. & Verma, I. M. (1979) *Cold Spring Harbor Symp. Quant. Biol.* **44**, in press.
11. Verma, I. M. (1978) *J. Virol.* **26**, 615–629.
12. Bolivar, F., Rodriguez, R. L., Green, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H. & Falkow, S. (1977) *Gene* **2**, 95–113.
13. Verma, I. M. & McKennett, M. A. (1978) *J. Virol.* **26**, 630–645.
14. Sutcliffe, J. G. (1978) *Nucleic Acids Res.* **5**, 2721–2728.
15. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
16. Sutcliffe, J. G. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77–90.
17. Gilboa, E., Goff, S., Shields, A., Yoshimura, F., Mitra, S. & Baltimore, D. (1979) *Cell* **16**, 863–874.

18. Shine, J., Czernilofsky, A. P., Friedrich, R., Bishop, J. M. & Goodman, H. M. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 1473–1477.
19. Stoll, E., Billeter, M. A., Palmenberg, A. & Weissman, C. (1977) *Cell* **12,** 57–72.
20. Van Beveren, C., Goddard, J. G., Berns, A. & Verma, I. M. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 3307–3311.
21. Rose, J., Haseltine, W. A. & Baltimore, D. (1976) *J. Virol.* **20,** 324–329.
22. Mitra, S. W., Goff, S., Gilboa, E. & Baltimore, D. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 4355–4359.
23. Kleckner, N. (1977) *Cell* **11,** 11–23.
24. Proudfoot, N. J. (1979) *Nature (London)* **297,** 376.
25. Sakonju, S., Bogenhagen, D. F. & Brown, D. D. (1980) *Cell* **19,** 13–25.
26. Sims, J. & Benz, E. W. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 900–904.
27. Chien, Y.-H., Verma, I. M., Shih, T. Y., Scolnick, E. M. & Davidson, N. (1978) *J. Virol.* **28,** 352–360.
28. Schultz, A. M. & Oroszlan, S. (1979) *Biochem. Biophys. Res. Commun.* **86,** 1206–1213.
29. Witte, O. N. & Wirth, D. (1979) *J. Virol.* **29,** 735–743.
30. Karshin, W. L., Arcement, L. J., Naso, R. B. & Arlinghaus, R. B. (1977) *J. Virol.* **23,** 787–798.
31. Hartley, J. W., Wolford, N. K., Old, L. H. & Rowe, W. P. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 789–793.
32. Donoghye, D. J., Rothenberg, E., Hopkins, S., Baltimore, D. & Sharp, P. A. (1978) *Cell* **14,** 959–970.
33. Elder, J. H., Gautsch, J. W., Jensen, F. C., Lerner, R. A., Hartley, J. W. & Rowe, W. P. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4676–4680.
34. Hicks, J. B., Strathern, J. N. & Herskowitz, I. (1977) in *DNA Insertion Elements, Plasmids, and Episomes,* eds. Bukhari, A., Shapiro, J. & Adhya, S. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 457–462.
35. Tonegawa, S., Maxam, A., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1485–1489.
36. Seidman, J. G., Leder, A., Nau, M., Norman, B. & Leder, P. (1978) *Science* **202,** 11–17.