Proc. Natl. Acad. Sci. USA
Vol. 77, No. 9, pp. 5234–5238, September 1980
Biochemistry

Nucleotide sequence analysis of the transforming region and large terminal redundancies of Moloney murine sarcoma virus

(*src* sequence/open reading frame/inverted repeats/transposable elements)

E. PREMKUMAR REDDY*, MARY JANE SMITH*, ELI CANAANI*, KEITH C. ROBBINS*, STEVEN R. TRONICK*, SAYEEDA ZAIN†, AND STUART A. AARONSON*

*Laboratory of Cellular and Molecular Biology, National Cancer Institute, Bethesda, Maryland 20205; and †Cancer Center, University of Rochester Medical School, Rochester, New York 14642

Communicated by Robert J. Huebner, June 26, 1980

ABSTRACT **The sequence of the transforming region of the Moloney murine sarcoma virus genome has been determined by using molecularly cloned viral DNA. This region, 3.6 to 5.8 kilobase pairs from the left end of the molecule, contains the entire cellular insertion (*src*) sequence as well as helper viral sequences including the large terminal repeat (LTR). On the viral RNA strand, a long (1224 bases) open reading frame commenced to the left of the *src*–helper virus junction and terminated at a point 58 nucleotides into helper viral sequences to the right of *src*. Possible promoter and acceptor splice signals were detected in helper viral sequences upstream from this open reading frame. On the antiviral RNA strand, several promoter-like sequences, including one within the *src* region itself, were identified. However, no open reading frame downstream from these promoters was detected in the antiviral RNA strand. The LTR was found to contain promoter-like sequences as well as mRNA capping and polyadenylylation signals. In addition, it possessed an 11-base inverted terminal repeat at each end. Thus, the structure of the Moloney murine sarcoma virus genome with an LTR at each end resembles that of prokaryotic transposable elements.**

Moloney murine sarcoma virus (M-MuSV) is a representative of the class of replication-defective retroviruses which transform fibroblasts in tissue culture and induce fibrosarcomas *in vivo*. Accumulating evidence indicates that this virus arose by recombination of the nondefective Moloney murine leukemia virus and cellular sequences (*src*) present within the normal mouse genome (1–6). By heteroduplex mapping (1, 2) and molecular hybridization techniques (3), the cell-derived *src* sequences have been localized to a specific region of the sarcoma viral genome.

The transforming region of the M-MuSV has been localized by comparison of restriction endonuclease maps of the parental virus and of deletion mutants which have lost up to 35% of the viral genome and yet retain transforming activity (4). Moreover, analysis of the transforming activity of subgenomic DNA fragments isolated after restriction endonuclease digestion of linear viral DNA has indicated that the *src* region is essential for transformation (4–6). The smallest fragment known to retain biological activity contained all *src* sequences and spanned the region 3.6 to 5.8 kilobase pairs (kbp) from the left end of the molecule (4).

With the advent of recombinant DNA techniques, it has become possible to clone sarcoma viral genes and thus to study their structural organization in greater detail. In an attempt to better understand the structural organization and possible molecular mechanisms involved in transformation by M-MuSV, we have undertaken the sequence analysis of the transforming region of the molecule. Putative regulatory signals for transcription, RNA processing, and translation of *src* sequences of M-MuSV have been identified. We have also sequenced the large terminal repeats (LTRs) of the viral DNA molecule. Inverted terminal repeats within LTRs suggest analogy between the MuSV genome and prokaryotic transposable elements.

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.

METHODS

Molecular Cloning. The Charon phage cloning system was provided by F. R. Blattner (7). The cloning of unintegrated circular M-MuSV-124 DNA in Charon 21A has been described in detail (8). For cloning of M-MuSV DNA in pBR322 plasmid (9), the viral insert was excised from M-MuSV DNA hybrid by cleavage with *Hin*dIII enzyme and then ligated to *Hin*dIII-digested pBR322. Hybrid DNA was transfected into $CaCl_2$-treated *Escherichia coli* C600 (10). Transformants were selected by plating on NZY agar (7) containing 40 μg of ampicillin per ml. All colonies were screened for the presence of MuSV DNA by the *in situ* hybridization method (11). Plasmid DNA was isolated from chloramphenicol-treated cells by standard procedures (12). The M-MuSV DNA subcloned in pBR322 was used for sequence analysis.

Nucleotide Sequencing. Restriction fragments of MuSV DNA cloned in pBR322 were obtained by sequential preparative agarose gel electrophoresis and DE-52 ion exchange chromatography (13). The fragments were labeled at their 5′ ends by using [γ-^{32}P]ATP and T4 polynucleotide kinase as described by Maxam and Gilbert (14). The fragments labeled at the two 5′ ends were further digested with appropriate restriction endonucleases and fractionated on agarose or polyacrylamide gels. These fragments, labeled only at one of the 5′ ends, were subjected to sequence determination by the procedures of Maxam and Gilbert (14) and Maat and Smith (15).

RESULTS

Sequencing Strategy. The covalently closed circular DNA of M-MuSV was recently cloned in λ phage (8). The 5.8-kbp DNA species, corresponding to the full-size genome of M-MuSV, was subcloned in pBR322 and used for sequence analysis. After isolation of the insert from pBR322 DNA, a detailed cleavage map (Fig. 1) of the transforming region of MuSV genome was developed by using the partial restriction mapping technique of Smith and Birnstiel (16). Fragments included in the transforming region were subjected to sequence determination by the partial chemical degradation method of Maxam and Gilbert (14) and by the nick-translation technique of Maat and Smith (15). Sequences of both strands were determined for most of the genome, and all restriction cleavage sites were

Abbreviations: M-MuSV, Moloney murine sarcoma virus; kbp, kilobase pair(s); LTR, large terminal repeat.

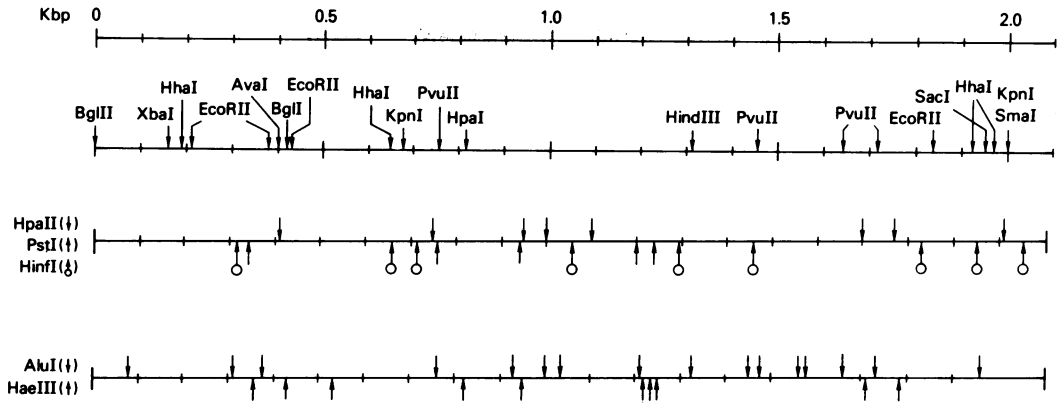

FIG. 1. Restriction enzyme map of the transforming region of M-MuSV. The restriction sites were determined by the partial restriction mapping technique of Smith and Birnstiel (16). Fragments generated by *Bgl* II, *Xba* I, *Bgl* I, *Hpa* I, *Hin*dIII, *Sma* I, *Hpa* II, *Hin*fI, and *Hae* III were used for sequence analysis.

confirmed by the sequence analysis. The entire nucleotide sequence is given in Fig. 2.

Molecular hybridization and heteroduplex analysis of M-MuSV (1, 6, 8) have demonstrated that within the transforming region there is a cellular insertion (*src*) sequence. By comparison of the sequence of the M-MuSV-transforming region reported here to the nucleotide sequence of Moloney murine leukemia virus (unpublished results) it was possible to localize the left junction between *src* and helper viral sequences to a point 26 nucleotides 3′ to the *Xba* I site and the right junction to a position 22 nucleotides 3′ to the *Hin*dIII site (Fig. 2).

Putative Promoter and Splicing Signals in the Transforming Region. Analysis of helper viral sequences to the right side of the *Bgl* II cleavage site revealed the occurrence of a transcription initiation signal (17, 18) at positions 29–35. Moreover, a dinucleotide 5′ A-G 3′ which is the common capping site of many eukaryotic mRNAs and is often found about 23 nucleotides from a transcriptional promoter signal (19) was found at positions 57–58 and 59–60, 22 and 24 nucleotides, respectively, downstream from the promoter-like sequence. In addition, a putative acceptor splice point was found at position 83. In general, splicing acceptor sites (at the 3′ end of intervening sequences) contain a pyrimidine-rich nucleotide track followed by the dinucleotide 5′ A-G 3′ (20). Indeed, the sequence A-G at 83–84 was preceded by a track of 11 pyrimidines lacking another A-G dinucleotide (20). Electron microscopic studies have shown that the M-MuSV genome retains the splice junction in Moloney murine leukemia virus at which 5′ leader sequences are joined to the body of the mRNA for the *env* gene (2, 21). The position of this splice point (2) closely corresponds to the position of the splicing signal identified here. This signal could serve a similar function in the processing of a putative *src* mRNA.

Sequence Organization of the M-MuSV *src* Region and Helper Viral Sequences Flanking Its Right Side. Examination of the viral RNA strand of the *src* region of MuSV (Fig. 1) revealed an open reading frame starting with the initiation codon ATG at position 176 and terminating with the triplet TAA at position 1400. This stretch of 1224 nucleotides began close to the *src*–helper viral junction on the left and ended 58 nucleotides into the helper viral sequences to the right of the *src* region. This segment of the genome has a coding capacity for a protein of 408 amino acids.

Toward the right end of the genome, we observed a cluster of A+T-rich sequences which included a stretch of eight base pairs at positions 1506–1513, followed by an inverted repeat of the same sequence at positions 1516–1523. The inverted repeats contained the polyadenylylation signal 5′ A-A-T-A-A-A 3′. This signal at positions 1507–1512 preceded the dinucleotide C-A in position 1530 by 18 base pairs, which is a preferred site for polyadenylylation (22).

The nucleotide sequence of the cDNA strand (antiviral RNA strand) included a promoter-like sequence (21), 5′ T-A-A-A-A-A-A-T 3′, at positions 1347–1340, 17 bases to the right of the *Hin*dIII site within the *src* sequence. Fifteen base pairs from this sequence in the 3′ direction was 5′ A-G 3′, a dinucleotide that is the common capping site (5′ end) of many eukaryotic mRNAs and is often found about 23 nucleotides from a transcriptional promoter signal (19). On the same strand immediately to the right of the promoter-like sequence described above were two similar sequences at positions 1358–1365 and 1371–1379, respectively. No open reading frame downstream from these promoter-like sequences could be identified.

Sequence of the LTR. One of the LTRs (23, 24) of M-MuSV extended 584 bases, between positions 1546 and 2127. We identified the extent of the repeat by additional sequence determination into the second LTR attached in tandem within the 5.8-kbp M-MuSV DNA (8). Examination of the LTR sequence revealed a number of salient features.

Inverted terminal repeats. An inverted repeat of 11 nucleotides, 5′ T-G-A-A-A-G-A-C-C-C-C 3′, appeared at the termini of the LTR (1546–1556, 2117–2127). Because genomic M-MuSV contains a complete LTR at each end, the entire M-MuSV genome possesses a terminal inverted repeat.

Identification of the 5′ end of M-MuSV RNA. Restriction enzyme analysis has indicated that the M-MuSV LTR is similar to that of the murine leukemia virus (25, 26). As such, it is composed of a track of about 450 nucleotides derived from the 3′ end of the viral RNA, directly followed to the right by a stretch of approximately 150 nucleotides derived from the 5′ end of the viral genome (27). To identify accurately these two sections on the terminal repeat, we compared the DNA sequence obtained here with certain previously reported murine leukemia virus sequences. The sequence 5′ A-A-T-G-A-A-A-G-A 3′ has been shown to reside at the 5′ terminus of "strong-stop" M-MuSV DNA (28) and, as such, to represent the region of joint between the latter and the tRNA primer. We found that the last seven nucleotides of M-MuSV LTR, 5′ T-G-A-A-A-G-A 3′, located at positions 2121–2127 on the anti-viral RNA strand corresponded to the sequence described above. We noticed the absence of the first two nucleotides, A-A, of strong-stop DNA in LTR. Similarly, these nucleotides are missing from the LTR of integrated M-MuSV (29). Having determined the position of the 5′ terminus of strong-stop DNA of LTR we expected the 5′ end of the viral RNA to reside around 135 nucleotides leftward on LTR (27, 28). Moreover, because the capped structure

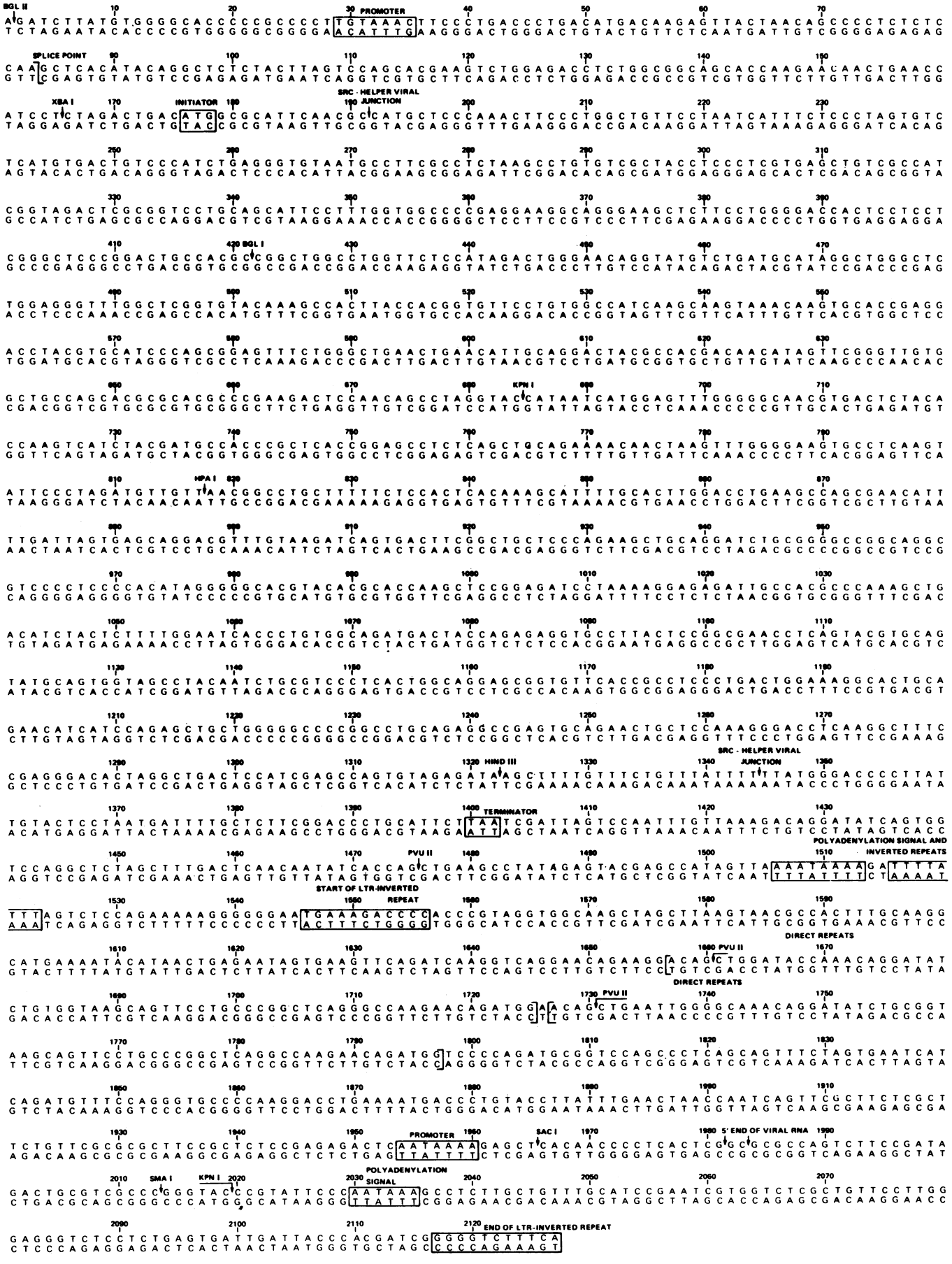

FIG. 2. Sequence of transforming region of M-MuSV DNA. The sequence described here encompasses the *src* region and the flanking helper viral sequences at each end of the *src* region. The top strand has the same polarity as genomic viral RNA. Unique restriction enzyme sites as well as some of the major structural features of the genome are indicated.

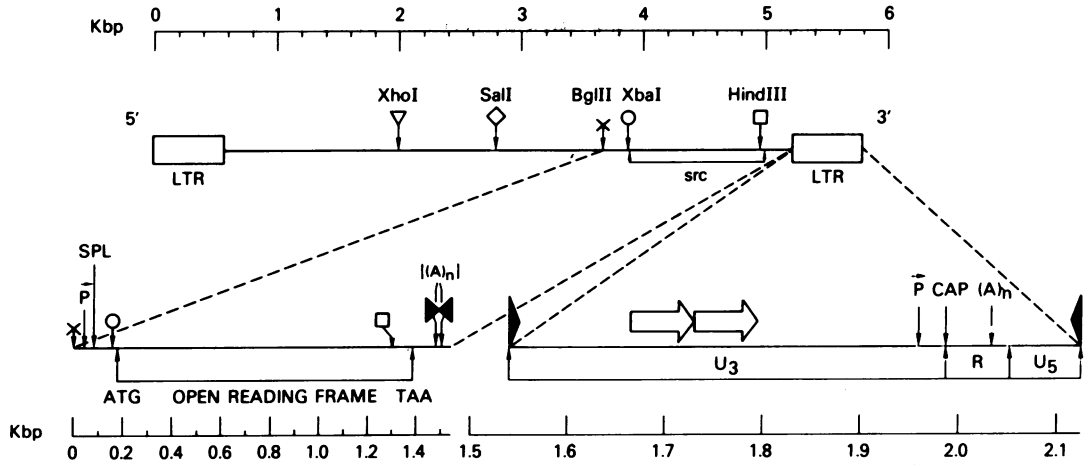

FIG. 3. Summary of the major structural features of the transforming region of M-MuSV. The upper diagram represents a physical map of the M-MuSV genome and indicates unique restriction enzyme cleavage sites. Sections of the transforming region have been expanded in the lower diagram and illustrate important features of the sequence. SPL, splicing signal; $\vec{P}$, promoter; $(A)_n$, polyadenylylation signal, CAP, 5′ terminus of M-MuSV genomic RNA; U_3, sequences unique to the 3′ end of genomic RNA; R, terminally redundant sequences of genomic RNA; U_5, sequences unique to 5′ end of genomic RNA; ⇒, direct repeat; ▶◀, inverted repeat.

at the extreme 5′ end of murine leukemia virus RNA has been shown to be meGpppGmeCG (30), the corresponding DNA sequence should be G-C-G. Either of the triplets at positions 1982–1984 and 1984–1986, localized at 145 and 143 nucleotides, respectively, from the 5′ end of strong-stop DNA were in appropriate positions to represent the capping sequence. Thus, either of these two triplets could correspond to the 5′ end of the viral RNA.

Transcription initiation and termination signals. A promoter-like sequence, 5′ A-A-T-A-A-A-A 3′, was found in the viral RNA strand at positions 1954–1960. This signal preceded by 22 and 24 nucleotides, respectively, the two GCG triplets likely to be at the 5′ end of the viral RNA. The correlation between the position of the promoter-like sequence and the 5′ end of the viral RNA strongly suggests that the sequence at positions 1954–1960 is the promoter for synthesis of M-MuSV genomic RNA. Similar to other promoter-like signals described above, this sequence also contained a polyadenylylation signal. Moreover, the dinucleotide C-A, which is a preferred polyadenylylation site, was found 16 nucleotides downstream from this sequence. An additional polyadenylylation signal, 5′ A-A-T-A-A-A 3′, was found at positions 2030–2035 and preceded C-A at position 2051–2052.

This polyadenylylation signal was included within a larger sequence at positions 1954–1959, which closely fit the sequence of oligonucleotide 21, found at both ends of murine leukemia virus RNA, and localized to the region very near the poly(A) tail (31). This correlation strongly suggests that the sequence at positions 2030–2035 represents the polyadenylylation signal for genomic viral RNA. It is known that genomic RNA contains a direct terminally repeated sequence (trs or R) of 50–60 nucleotides (31). Thus, if the C-A signal at position 2051–2052 corresponds to the 3′ end of M-MuSV RNA, the sequence of 68–70 nucleotides between this C-A and the GCG triplets representing the 5′ end of the viral RNA (at positions 1982–1984 or 1984–1986) should constitute the R region of MuSV.

Sequence duplication. M-MuSV LTR was found to contain a nearly perfect duplication of 72 or 73 bases at positions 1657–1725 and 1726–1797. The function of these sequences is not known.

The sequence of the LTR of the integrated ml-MuSV proviral DNA has been independently determined recently by Dhar *et al.* (29). That sequence is similar to the DNA sequence of the LTR of unintegrated M-MuSV sequences described here.

DISCUSSION

Nucleotide sequence analysis of the region of the M-MuSV genome corresponding to the smallest known fragment retaining transforming activity (4) has revealed several important features of its molecular organization. The M-MuSV transforming region, which extends from the *Bgl* II site to the right end of the genome (4), is composed of 191 bases of helper viral sequences followed by the entire cellular insertion (*src*) sequence and ending with 783 bases of helper viral information including the LTR. By comparison of the nucleotide sequence of the M-MuSV transforming region with information available from sequence studies of Moloney murine leukemia virus, it was possible to localize junction points between *src* and helper viral sequences.

Examination of the sequence of the transforming region revealed a single open reading frame on the viral RNA strand. This frame, 1224 bases long, commenced to the left of the *src*–helper viral junction and terminated at a point 58 nucleotides into the helper viral sequences to the right of *src* (Fig. 3). Our findings imply that, if this open reading frame were the coding sequence for the M-MuSV transforming protein, the first 16 bases which include the initiator codon and the last 58 bases including the terminator codon are contributed by the helper viral genome. This sequence could code for a protein of 408 amino acids and a molecular weight of around 49,000.

A promotor-like sequence could be identified at positions 29–36 within the transforming region and thus could serve as an initiation signal for transcription of the *src* gene. Alternatively, the promoter-like sequence within the left LTR could be the normal initiation signal for *src* gene transcription. If the latter were the case, the acceptor splice point located between *Bgl* II and *Xba* I sites could be used for splicing of a leader sequence initiated near the left LTR promoter to a *src* transcript. Transcription of the *Bgl* II transforming fragment would then presumably require integration of the fragment near a cellular transcription signal or might involve rearrangement of the promoter within the LTR at the 3′ end of the genome to a position upstream from the *src* gene.

Several promoter-like sequences, one of which was included within *src*, were detected to the right of the *Hin*dIII site on the antiviral RNA strand. However, numerous terminator codons were found in each of the reading frames downstream from these promoters. Thus, even if the negative-strand RNA were synthesized, it is unlikely to code for any protein. Further

knowledge of the transcriptional and translational products of the M-MuSV transforming gene will be required in order to ascertain whether the antiviral RNA strand plays any role in transformation.

The LTRs that occur at both 5′ and 3′ ends of the integrated viral genome had several unique structural features. They contained both promoter-like sequences for the initiation of viral RNA synthesis and signals for the polyadenylylation of viral mRNAs. The promoter-like sequence at position 1954–1960 preceded by 22 or 24 nucleotides the proposed 5′ end of the viral RNA. This sequence was followed 70 bases later by the polyadenylylation signal. The positioning of LTRs at both ends of the integrated provirus raises the possibility that the genomic viral RNA is initiated at the promoter of the left LTR and terminated at the polyadenylylation signal of the right LTR. The juxtaposition of promoter and termination signals within the LTR might also result in the formation of short RNA transcripts or in transcripts initiating or terminating in flanking cellular DNA. On the other hand, there may exist a mechanism for prevention of premature termination at the polyadenylylation site located 70 bases downstream from the promoter in the left LTR.

The finding of an inverted repeat sequence of 11 nucleotides at the termini of each LTR emphasizes some significant similarities between retroviruses and prokaryotic transposable elements (32, 33). Because of the occurrence of the LTR sequences at both ends of the integrated viral genome, the total M-MuSV genome has an 11-base-pairs inverted repeat at the two ends of the molecule. Thus, the M-MuSV genome resembles prokaryotic transposable elements which also are flanked by short inverted repeats and are capable of translocation to different positions on the chromosome or to another replicon in the cell. Like these bacterial elements, retroviruses integrate into the host DNA in a linear orientation with defined end points. Finally, as in the case of transposable elements such as Tn*10* (34, 35), we have recently found that M-MuSV DNA is capable of promoting deletion events originating at the very end of the LTR (unpublished data). It is possible that, similar to integrated prokaryotic transposable elements which cause strong polar effects and rearrangements of the host DNA, when the genomes of retroviruses are integrated at the appropriate position they may directly affect the expression of neighboring host genes. In contrast to replication-defective retroviruses such as M-MuSV, replication competent retroviruses that cause leukemia as yet have not been shown to possess discrete transforming genes. The oncogenicity of these latter viruses could be related to their transposon-like structure.

1. Hu, S., Davidson, N. & Verma, I. M. (1977) *Cell* **10**, 469–477.
2. Donoghue, D. J., Sharp, P. A. & Weinberg, R. A. (1979) *Cell* **17**, 53–63.
3. Dina, D. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2694–2698.
4. Canaani, E., Robbins, K. C. & Aaronson, S. A. (1979) *Nature (London)* **282**, 378–383.
5. Andersson, P., Goldfarb, M. P. & Wienberg, R. A. (1979) *Cell* **16**, 63–75.
6. Oskarsson, M., McClements, W. L., Blair, D. G., Maizel, J. V. & Vande Woude, G. F. (1980) *Science* **207**, 1222–1224.
7. Helvet, L. R., Daniels, D. L., Schzoeder, L. Z., Williams, B. G., Denniston-Thompson, K., Moore, D. D. & Blattner, F. R. (1980) *J. Virpl.* **33**, 401–410.
8. Tronick, S. R., Robbins, K. C., Canaani, E., Devare, S. G., & Aaronson, S. A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6314–6318.
9. Bolivar, F., Rodriguez, R. L., Bethlach, M. & Boyer, H. W. (1977) *Gene* **2**, 75–93.
10. Taketo, A. (1972) *J. Biochem.* **72**, 973–979.
11. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–181.
12. Biznboin, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513–1523.
13. Zain, S. & Roberts, R. J. (1979) *J. Mol. Biol.* **131**, 341–352.
14. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
15. Maat, J. & Smith, A. J. H. (1978) *Nucleic Acids Res.* **5**, 4537–4545.
16. Smith, H. O. & Birntstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2399.
17. Pribnow, D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 784–788.
18. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
19. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
20. Seif, I., Khoury, G. & Dhar, R. (1979) *Nucleic Acids Res.* **6**, 3387–3398.
21. Rothenberg, E., Donoghue, D. J. & Baltimore, D. (1978) *Cell* **13**, 435–451.
22. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214.
23. Hsu, T. W., Salzen, J. L., Mark, G. E., Guntaka, R. V. & Taylor, J. M. (1978) *J. Virol.* **28**, 810–818.
24. Shank, P. R., Hughes, S. H., Kung, J. H., Majors, J. E., Quintrel, N., Guntaka, R. V., Bishop, J. M. & Varmus, H. E. (1978) *Cell* **15**, 1383–1396.
25. Gilboa, E., Goff, S., Shields, A., Yoshimura, F., Mitra, S. & Baltimore, D. (1979) *Cell* **16**, 863–874.
26. Benz, E. W. & Dina, D. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3294–3298.
27. Gilboa, E., Mitra, S., Goff, S. & Baltimore, D. (1979) *Cell* **18**, 93–100.
28. Haseltine, W. A., Kleid, D. G., Panet, A., Rothenberg, E. & Baltimore, D. (1976) *J. Mol. Biol.* **106**, 109–131.
29. Dhar, R., McClements, W. L., Enquist, L. W. & Vande Woude, G. W. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3937–3941.
30. Rose, J. K., Haseltine, W. A. & Baltimore, D. (1976) *J. Virol.* **20**, 324–329.
31. Coffin, J. M., Hageman, T. C., Maxam, A. M. & Haseltine, W. A. (1978) *Cell* **13**, 761–773.
32. Bukhari, A. J., Shapiro, J. A. & Adhya, S. L., eds. (1977) *DNA Insertion Elements, Plasmids and Episomes* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
33. Ohtsubo, H., Ohmori, H. & Ohtsubo, E. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **53**, 1269–1277.
34. Ross, D. G., Swan, J. & Kleckner, N. (1979) *Cell* **16**, 721–731.
35. Starlinger, P. (1980) *Plasmid* **3**, 241–259.