Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 7041–7045, October 1987
Biochemistry

Characterization of mouse mammary tumor virus *gag-pro* gene products and the ribosomal frameshift site by protein sequencing

A. HIZI*, L. E. HENDERSON, T. D. COPELAND, R. C. SOWDER, C. V. HIXSON, AND S. OROSZLAN†

Laboratory of Molecular Virology and Carcinogenesis, Bionetics Research, Inc.–Basic Research Program, National Cancer Institute–Frederick Cancer Research Facility, Frederick, MD 21701

Communicated by Harold E. Varmus, June 26, 1987 (received for review April 23, 1987)

ABSTRACT **The synthesis of retroviral polyproteins that are the translational products of the genome-size mRNA is initiated in the upstream *gag* gene. The synthesis of the products of the protease gene (*pro*) and polymerase gene (*pol*) is regulated by translational suppression (in-frame read-through or frameshift) of termination codons as a strategy developed for controlling the level of replicative enzymes required only in catalytic amounts. In mouse mammary tumor virus (MMTV), three overlapping reading frames are utilized for the synthesis of *gag*-encoded Pr77, *gag-pro*-encoded Pr110, and *gag-pro-pol*-encoded Pr160 polyproteins. To characterize *pro* gene products and to determine the site of frameshift required for the synthesis of Pr110, we purified and sequenced three MMTV proteins: p14, p30, and p13. Sequence analysis showed that p14 is the basic nucleic acid-binding protein derived entirely from *gag*, and p13 is a product of the *pro* gene and has characteristic sequences of protease. A comparison of the amino acid sequences of p30 with the corresponding nucleotide sequence of proviral DNA allowed the delineation of the frameshift site utilized *in vivo* for the synthesis of the *gag-pro*-encoded fusion polyprotein Pr110. The results showed that: (*i*) the N-terminal 94 residues of p30 are translated from the *gag* frame; (*ii*) residue leucine-95 is specified by either the last UUG codon of *gag* or the overlapping CUU codon in the *pro* frame; and (*iii*) the elongation of the peptide chain from residue 96 continued to be encoded in the *pro* frame to the *pro* terminator. The possible mechanisms of frameshift and of the tRNAs involved are discussed.**

The genome of the replication-competent retroviruses has most commonly been characterized as a positive-strand RNA of 8.5–9.0 kilobases (kb) that is divided into three genes arranged in the order of 5′ *gag-pol-env* 3′ (1). While the protein products of the *env* gene are translated from a subgenomic (4.0 kb) spliced mRNA, the protein products of *gag* and *gag-pol* are synthesized via the translation of the genome-size mRNA. Although the *gag*- and *gag-pol*-encoded polyproteins of all known retroviruses are initiated at the same AUG codon of 5′ *gag*, the gag-pol precursor is generally expressed at only about 5–20% of the level of the gag precursor. Further, in mammalian type C viruses, the *gag* and *pol* gene products are encoded in the same reading frame, but for all other retroviruses, the *gag* and *pol* reading frames are out of phase, either overlapping each other or separated by a third reading frame that overlaps both (for review, see ref. 1). Recent investigations led to the structural and functional characterization of viral components encoded between the 3′ end of the *gag* gene and the codon that specifies the N terminus of reverse transcriptase (RT) in the *pol* gene. Such studies opened ways to elucidating the mechanisms of translational regulation of retroviral protein expression. Thus, reports from this laboratory (2, 3) established that the *gag* amber terminator of both murine and feline leukemia viruses (MLV and FeLV, respectively) is suppressed, resulting in in-frame read-through for the synthesis of protease–reverse transcriptase–endonuclease of the *gag-pol*-encoded polyprotein. Furthermore, Jacks and Varmus (4), who utilized RNA transcripts of wild-type and mutant DNAs of Rous sarcoma virus in a eukaryotic cell-free translational system, demonstrated convincingly that the *gag* termination codon of Rous sarcoma virus (5) can be bypassed by a mechanism that involves a ribosomal frameshift event.

We have been involved in chemically characterizing the proteins of mouse mammary tumor virus (MMTV) which is a prototype type B retrovirus. As shown previously by others, the *gag* gene of MMTV is translated into the gag precursor polyprotein Pr77 having a map order of NH_2-p10-pp21-p27-p14-OH and into two larger *gag*-related precursors, Pr110 and Pr160 containing additional sequences (reviewed in ref. 6). A minor MMTV protein, p30, was shown to contain tryptic peptides of p14 derived from Pr77 and additional peptides shared with Pr110 (6). We have now purified and partially sequenced all the proteins from MMTV accounting for Pr77 and Pr110. While these studies were in progress the relevant nucleotide sequences of the MMTV genome were determined, and the results indicated the presence of three out-of-phase overlapping reading frames: one for *gag*, one designated *pro* for protease, and one for *pol* (7, 8). In this communication we report the results of our protein sequencing studies aimed at determining the translational frameshift site required for the synthesis of Pr110. The combined protein and DNA sequence data clearly show that p30 is synthesized from a portion of the genome spanning the *gag-pro* junction via a translational frameshift in the −1 direction. We also report the identification and analysis of a p13 protein derived from the 3′ end of $Pr110^{gag\text{-}pro}$.

MATERIALS AND METHODS

MMTV (C3H strain) grown in Mm5mt cells was obtained from the Biological Products Laboratory (Program Resources, Frederick Cancer Research Facility), and the R-III strain was from the repository of the National Cancer Institute. Viral proteins were purified by reversed-phase HPLC (RP-HPLC) on a μBondapak C_{18} (19 × 150 mm) column (Waters) by methods previously described (9). For proteolytic digestion, purified protein (150 μg) was dissolved in 0.1 ml of 0.01 M sodium acetate (pH 4.0) and heated to 100°C for 10 min, cooled, added to 3 μg of V8 protease from *Staphylococcus aureus* (Sigma), and kept at room temperature for 24 hr. The resulting peptides were purified by RP-HPLC on a μBondapak C_{18} (2.0 × 300 mm) column (9).

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: MMTV, mouse mammary tumor virus; FeLV, feline leukemia virus; BLV, bovine leukemia virus; MLV, murine leukemia virus; RP-HPLC, reversed-phase HPLC.

*Permanent address: Sackler School of Medicine, Tel Aviv University, Israel.
†To whom reprint requests should be addressed.

N-terminal amino acid sequences were determined by automated Edman degradation in a gas-phase sequencer (Applied Biosystems, Foster City, CA) as described (10). C-terminal amino acid sequences were determined by carboxypeptidase A digestion as described (11).

RESULTS

The proteins of MMTV were dissociated from sucrose density gradient-purified virus by guanidine hydrochloride under reducing conditions and were purified by RP-HPLC. A typical elution profile from the RP-HPLC column is shown in Fig. 1. The proteins separated by this high-resolution method were identified by subsequent chemical analysis; in the order of their elution from the column, they are: p14, the basic nucleic acid-binding protein; pp21, a phosphorylated core protein; p13, a newly identified protein (see below); p30, shown previously to share some peptides with p14 (6); gp52, the surface glycoprotein; p27, the major core shell protein; p10, the myristylated N-terminal polypeptide of Pr77 (12); and gp36, the transmembrane protein. The chemical characterization of the glycosylated envelope proteins has been reported (13). The chemical characterization of the *gag* gene products p10, pp21, p27, and p14 will be described in detail elsewhere. The proteins of major interest for this study were p14, p30, and p13. Each of these proteins was rechromatographed by RP-HPLC and analyzed by sodium dodecyl sulfate/polyacrylamide gel electrophoresis as shown in Fig. 2. Purified p14, p30, and p13 were each analyzed by Edman degradation and by carboxypeptidase A digestion for the C-terminal amino acid sequence. As expected, p14 was found to be fully encoded by a 3′-end portion of the *gag* open reading frame, which we call "*gag-orf*." The N-terminal amino acid sequence of p13 was found to be:

Trp-Val-Gln-Glu-Ile-Ser-Asp-.

This sequence perfectly matched that predicted from the nucleotide sequence in the second open reading frame called "*pro-orf*." The N terminus tryptophan of p13 is located 115 codons upstream from the *pro* terminator (7, 8). The C terminus of p13 was determined to be leucine in agreement with the last codon in *pro-orf*. These data clearly indicate that the entire p13 is derived from the 3′ region of *pro-orf*, which, as previously noted by Jacks *et al.* (7) and Moore *et al.* (8), has the two highly conserved regions of retroviral proteases (2, 15). Our analysis now shows that both of these protease-specific sequences are entirely included in p13.

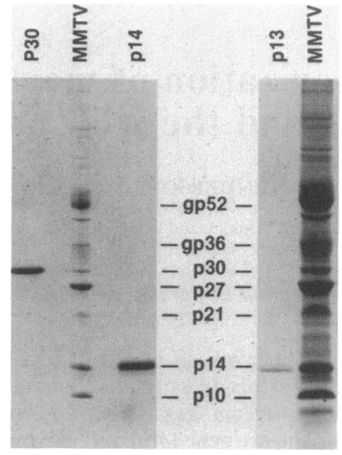

FIG. 2. MMTV p14, p30, and p13 (Fig. 1) were rechromatographed by RP-HPLC (not shown) and analyzed by sodium dodecyl sulfate/polyacrylamide gel electrophoresis (14).

We determined that the nucleic acid-binding protein, p14 and p30 have identical N-terminal amino acid sequences:

Ala-Ala-Ala-Met-Arg-Gly-Gln-Lys-Try-Ser-Thr-Phe-.

Their common N terminus was localized in the translated nucleotide sequence of the MMTV *gag-orf* 95 codons upstream from the *gag* terminator (7, 8). The determined C-terminal sequence of p14, -Lys-Asn-Leu-OH, was also found to be identical to the sequence predicted by the last three codons of MMTV *gag-orf*. The determined C-terminal sequence of p30, -Val-His-OH, was located in *pro-orf* 153–154 codons downstream from the *gag* terminator. The C terminus (histidine) of p30 is a residue located adjacent to the N terminus (tryptophan) of p13. These results showed that, in addition to p14, the N-terminal portion of p30 is also a translational product of *gag-orf*, but the C-terminal portion of the p30 protein is a product of *pro-orf*—i.e., p30 is a transframe protein.

In order to define the site of frameshift, the purified transframe protein p30 was digested with V8 protease, and the resulting peptides were separated by RP-HPLC as shown in Fig. 3. Each purified peptide was analyzed by amino acid analysis and Edman degradation, and the results were compared to the translated nucleotide sequence to locate peptides in the predicted amino acid sequence of p30 (7, 8). These results define the structure of p30 as shown in Fig. 4.

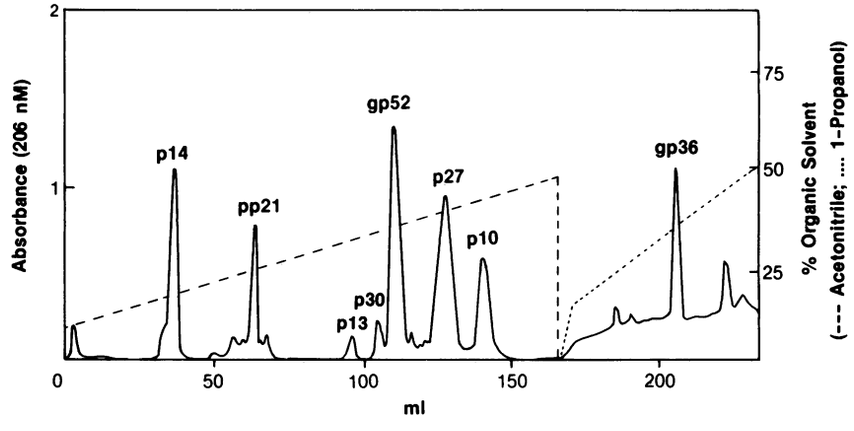

FIG. 1. Separation of MMTV proteins by RP-HPLC. Whole virus (50 mg) was disrupted in 50 ml of saturated guanidine hydrochloride containing 2% 2-mercaptoethanol, made pH 2 with the addition of trifluoroacetic acid, and applied onto a μBondapak C_{18} column (19 × 150 mm). Gradient elution was accomplished at pH 2 with an increasing concentration of acetonitrile in 0.05% trifluoroacetic acid (flow rate, 5 ml/min), and proteins were detected by UV absorption at 206 nm.

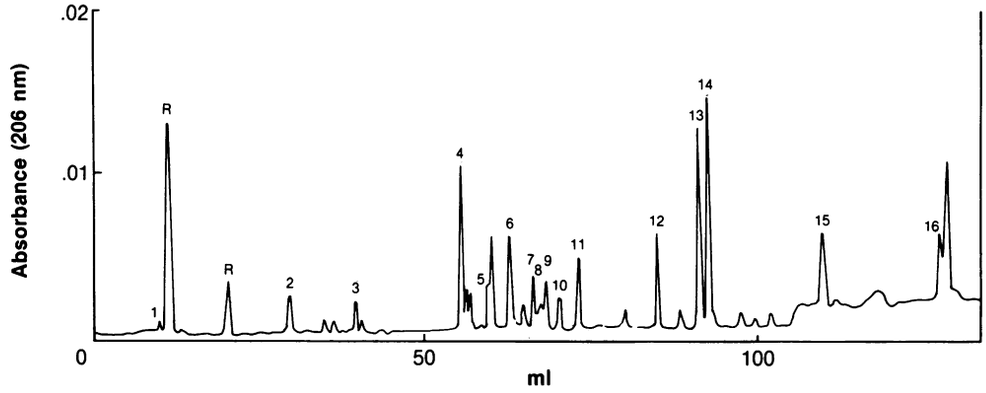

FIG. 3. Purified p30 (150 μg) was digested with staphylococcal V8 protease (Sigma), and the peptides were purified by RP-HPLC on a 2.0 × 30 mm column at 1.0 ml/min as in Fig. 1. Peptides of p30 were separated with a linear gradient of 0–50% acetonitrile over 2 hr. Peptides were detected by UV absorption at 206 nm and collected, and solvents were removed by lyophilization. Each peptide was analyzed for total amino acid content after acid hydrolysis and for N-terminal amino acid sequence by Edman degradation in a gas-phase sequencer. The determined amino acid sequences and composition were compared to the translated DNA sequence as shown in Fig. 4. A list of the numbered UV peaks (peptides) and the determined peptide segment of p30 is given below. Peptide segments are denoted by the residue number (Fig. 4) of the first and last residue of the peptide determined by Edman degradation (not necessarily the last residue of the peptide). Peaks: 1, residues 244–249; 2, 71–78; 3, 237–249; 4, 48–72; 5, 79–90; 6, 167–181; 7, 1–25; 8, 1–20; 9, 1–15; 10, 1–24; 11, 185–198; 12, 151–166; 13, 91–121; 14, 79–101; 15, 198–226; 16, 151–189. The peak labeled R contained reagents but no peptide.

The amino acid sequence spanning the *gag-pro* junction was determined from the analyses of peptides 13 and 14 (Fig. 3). Peptide 14 was a product of incomplete proteolysis beginning at residue 78 and overlapped peptide 13 beginning

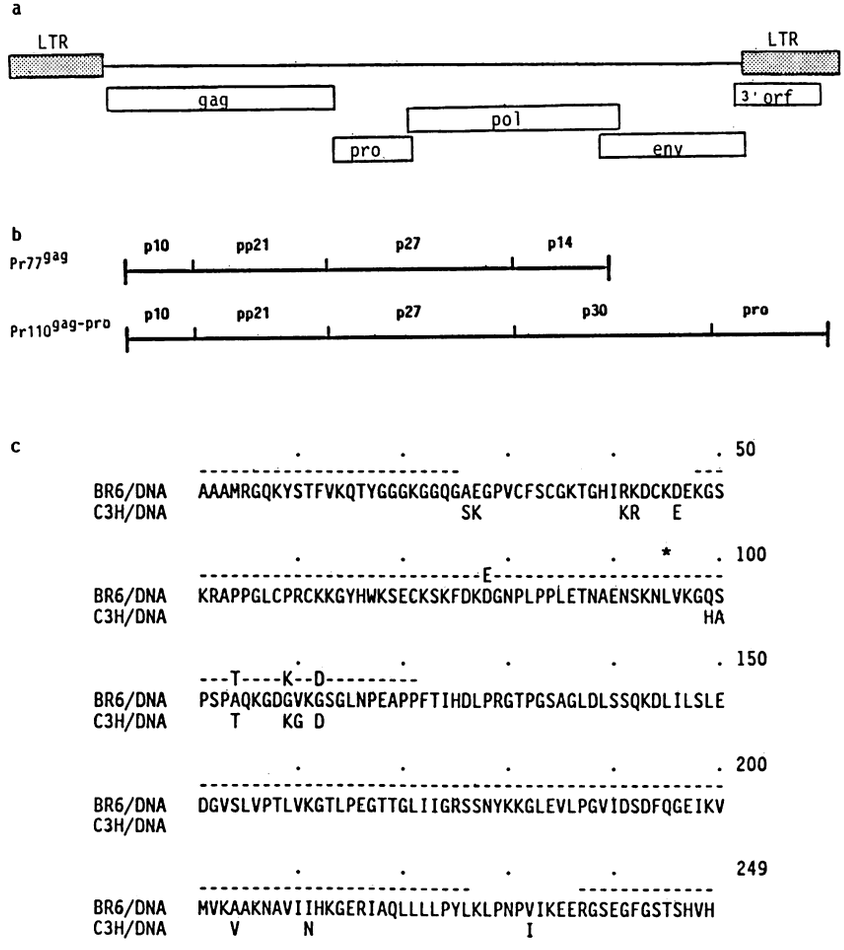

FIG. 4. (*a*) Genome organization and disposition of open reading frames of MMTV provirus (7, 8). (*b*) Map order of the primary translational products of the *gag* and *gag-pro* regions, the first two open reading frames from which the p30 transframe protein is derived. (*c*) The amino acid sequence of MMTV p30 predicted by translation of the DNA sequence of the BR6 strain (8) shown in single-letter amino acid code. Residues predicted for the C3H strain (7) are shown only where the predicted sequence of the BR6 strain differs from the C3H strain. Overlined residues were determined by Edman degradation of the MMTV R-III protein and peptides produced by V8 protease cleavage (see Fig. 3). Where the determined amino acid sequence of the protein differs from the translated BR6/DNA sequence, the determined residue is indicated above the predicted residue. The last residue that may be coded by *gag-orf* for p14 (leucine-95) is marked with an asterisk (*).

```
DNA      G C T  G A A  A A T  T C A  A A A  A A C  T T G  T A A  A G G  G G C  A G T  C C C  C T
gag-orf  Ala    Glu    Asn    Ser    Lys    Asn    Leu    ***

p30      ALA    GLU    ASN    SER    LYS    ASN    LEU    VAL    LYS    GLY    GLN    SER    PRO

pro-orf         ***    Lys    Phe    Lys    Lys    Leu    Val    Lys    Gly    Gln    Ser    Pro
DNA      G C  T G A  A A A  T T C  A A A  A A A  C T T  G T A  A A G  G G G  C A G  T C C  C C T
```

FIG. 5. Comparison of the nucleotide sequence of the segment of MMTV (BR6) DNA spanning the *gag-pro* overlap window with the corresponding amino acid sequence of transframe protein p30. Top *gag-orf* (0 frame); bottom *pro-orf* (−1 frame).

at residue 91 (Fig. 4). In Fig. 5, a portion of the p30 amino acid sequence (Ala-89 through Pro-101) is aligned with the nucleotide sequence spanning the *gag-pro* overlap window and amino acid sequences predicted in both *gag-orf* and *pro-orf*. The alignment shows that during protein synthesis, the frameshift event occurred just before or when elongation had reached the last codon in *gag-orf*. The leucine residue at position 95 of p30 (Fig. 5) could have been encoded by either the TTG codon ("0" frame) in *gag-orf* or the CTT codon in *pro-orf* (−1 frame). The sequence continues in the transframe protein from Val-96 to Pro-101 as shown in Fig. 5 and beyond to the C-terminal residue His-249 (see Fig. 4) in *pro-orf*.

DISCUSSION

We report in this communication the isolation and partial sequence analysis of three MMTV-coded proteins: p14, which is derived from *gag*; p30, which is the transframe protein derived from *gag* and 5′ *pro-orf*; and p13, which is the putative protease derived from 3′ *pro-orf*, representing the C-terminal portion of polyprotein Pr110. This together with the previously determined order of the proteolytic cleavage products in Pr77 *gag* defines the map order of Pr110 to be NH_2-p10-pp21-p27-p30-p13-OH. p30 and p13 seem to accumulate in the MMTV in near equimolar amounts but each is at a much lower level than p14 (Fig. 1). The finding that p30 contains the complete amino acid sequence of the p14 nucleic acid binding protein in its N-terminal region makes it likely that p30 also possesses nucleic acid-binding properties. Furthermore, the fact that a long C-terminal segment of 154 amino acids of p30 has high homology with the predicted sequence of type D viruses (16, 17) strongly suggests a common biological role for these sequences in these two divergent groups of retroviruses. A possible role in the formation of intracytoplasmic A particles has been proposed (7). The MMTV p30 structure is unique in that it is the only known transframe protein at the junction of the *gag* and *pro* reading frames in retroviruses that contains the entire nucleic acid binding protein domain plus long additional sequences.

MMTV p13 shows sequence homology to other retroviral proteases but as of this writing we do not have any evidence for the enzymatic activity of this protein. One might assume that a minor related protein product (p13′) that would extend some amino acids beyond the *pro* stop codon into the adjacent *pol* frame could be the active protease. Two frameshift events have been shown to be required for the synthesis of the *gag-pro-pol*-encoded polyprotein Pr160 (7, 8). In this study we have characterized the first transframe product, p30, of the *gag-pro* junction. The second transframe product spanning the *pro-pol* junction remains to be identified and characterized. Nevertheless, the protein data presented here clearly indicate that the translational product of the relatively long *pro-orf* of MMTV is processed into two smaller proteins, p30 and p13. Thus, the putative protease of MMTV (either p13 or p13′) is likely to be similar in size to that of MLV, FeLV, and BLV proteases, each consisting of 125 amino acids (2, 3, 18). A long *pro-orf* was also noted in type D viruses (16, 17). It is expected that, like in MMTV, proteolytic cleavage of type D virus gag-pro polyprotein will also yield protease molecules smaller than those predicted based on the length of *pro-orf* (16, 17).

Although the *gag-pro* overlap window and the translational frameshift event required for the synthesis of MMTV Pr110 were defined by DNA sequence analysis coupled with *in vitro* transcription and translation experiments (7, 8), direct evidence for the location of the frameshift site used under natural conditions was obtained by sequence analysis of the virion protein as described in this report. The protein data presented here confirm a frameshift event in the −1 reading frame, but the exact mechanism of frameshifting remains unclear. Due to the given codons specified in the zero and −1 reading frames, two alternatives exist (see Fig. 5). In the first one, the frameshift occurs between the UUG leucine codon and the UAA termination codon by reading the G twice, once in the UUG (0 frame) and then in the GUA valine codon (−1 frame). The second possibility is a shift 3 bases upstream where the C is read twice, once in the AAC triplet (0 frame) as asparagine and then in CUU in the −1 frame as leucine (Fig. 5). The second alternative is consistent with the involvement of the mainly homopolymeric AAAAAAC sequence identified by Jacks *et al.* (7) as the "frameshift signal" in the MMTV *gag-pro* overlap. This sequence also was predicted as a potential frameshift site in BLV and has parallels in prokaryotic systems (18–20). The simplest model involving such a signal sequence as proposed by Jacks *et al.* (7) "would call for tRNA reading the 0 frame codons to slip back one nucleotide and pair with the codon in the −1 frame." Based on the protein data, the first codon possibly read in this frame by a cognate tRNA ($tRNA^{Leu}$) is the CUU triplet. To allow this to occur, the $tRNA^{Asn}$ is required to slip back to decode the AAA triplet as asparagine and restore the reading frame. Another mechanism could involve recognition of a 0 frame codon by a noncognate normal tRNA utilizing only two of the three possible base pairs, as has been observed in a cell-free prokaryotic translational system (21). The protein data indicate that, in such a case, a normal $tRNA^{Leu}$ having AAG in positions 34–36 in the anticodon loop could promote a two-base translocation at the UUG leucine codon with a resultant shift into the −1 frame. In any case, additional factors must operate to confer specificity to the site of the frameshift event. The role of surrounding sequences and of a potential stem-loop structure in the viral RNA downstream to the overlap window has been discussed in some detail by Jacks *et al.* (7).

Further studies to determine the site of the −1 frameshift event in other retroviruses will be needed to help elucidate factors involved in determining the specificity for the event. Very recently a natural suppressor glutamine tRNA has been shown to occur at highly elevated levels in mouse cells infected with Moloney-MLV (22) and has been suggested to be responsible for the in-frame read-through between the *gag* and *pol* genes (2). The role of suppressor tRNAs in eukaryotic ribosomal frameshifting should also be considered. Retroviruses and retroviral-like elements in yeast (23, 24) and *Drosophila* (25) provide extremely useful system(s) to study the regulation of translation of eukaryotic mRNAs. Identification of all the important factors involved may help find specific ways for therapy of retroviral infections.

We thank Young Kim for excellent technical assistance. We are grateful to Harold Varmus, John Majors, Gordon Peters, and Clive Dixon for communicating results prior to publication. We also thank

Dolph Hatfield and Alan Rein for valuable discussions and Cheri Rhoderick and Carolyn Phillips for the preparation of the manuscript. This work was sponsored by the National Cancer Institute under Contract NO1-CO-23909 with Bionetics Research, Inc.

1. Weiss, R., Teich, N., Varmus, H. & Coffin, J., eds. (1982) *RNA Tumor Viruses* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
2. Yoshinaka, Y., Katoh, I., Copeland, T. D. & Oroszlan, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1618–1622.
3. Yoshinaka, Y., Katoh, I., Copeland, T. D. & Oroszlan, S. (1985) *J. Virol.* **55,** 870–873.
4. Jacks, T. & Varmus, H. E. (1985) *Science* **85,** 1237–1242.
5. Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) *Cell* **32,** 853–869.
6. Dickson, C. & Peters, G. (1983) *Curr. Top. Microbiol. Immunol.* **106,** 1–34.
7. Jacks, T., Townsley, K., Varmus, H. E. & Majors, J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4298–4302.
8. Moore, R., Dixon, M., Smith, R., Peters, G. & Dickson, C. (1987) *J. Virol.* **61,** 480–490.
9. Henderson, L. E., Sowder, R., Copeland, T. D., Smythers, G. & Oroszlan, S. (1984) *J. Virol.* **52,** 492–500.
10. Hewick, R. M., Hunkapiller, M. M., Hood, L. E. & Dryer, W. J. (1981) *J. Biol. Chem.* **256,** 7990–7997.
11. Oroszlan, S., Henderson, L. E., Stephenson, J. R., Copeland, T. D., Long, C. W., Ihle, J. N. & Gilden, R. V. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1404–1408.
12. Schultz, A. M. & Oroszlan, S. (1983) *J. Virol.* **46,** 355–361.
13. Henderson, L. E., Sowder, R., Smythers, G. W. & Oroszlan, S. (1983) *J. Virol.* **48,** 314–319.
14. Laemmli, U. K. (1970) *Nature (London)* **227,** 680–685.
15. Toh, H., Ono, M., Saigo, K. & Miyata, T. (1985) *Nature (London)* **315,** 691.
16. Power, M., Marx, P., Bryant, M., Gardner, M., Barr, P. & Luciw, P. (1986) *Science* **231,** 1567–1572.
17. Sonigo, P., Barker, C., Hunter, E. & Wain-Hobson, S. (1986) *Cell* **45,** 375–385.
18. Yoshinaka, Y., Katoh, I., Copeland, T. D., Smythers, G. W. & Oroszlan, S. (1986) *J. Virol.* **57,** 826–832.
19. Rice, N. R., Stephens, R. M., Burny, A. & Gilden, R. V. (1985) *Virology* **142,** 357–377.
20. Weiss, R. B. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5797–5801.
21. Dayhuff, T. J., Atkins, J. F. & Gesteland, R. F. (1986) *J. Biol. Chem.* **261,** 7491–7500.
22. Kuchino, Y., Beier, H., Akita, N. & Nishimura, S. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2668–2672.
23. Clare, J. & Farabaugh, P. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2829–2833.
24. Mellor, J., Malim, M. H., Gull, K., Tuite, M. F., McCready, S., Dibbayawan, T., Kingsman, S. M. & Kingsman, A. J. (1985) *Nature (London)* **318,** 583–586.
25. Saigo, K., Kugimiya, W., Matsuo, Y., Inouye, S., Yoshioka, K. & Yuki, S. (1984) *Nature (London)* **312,** 659–661.