Isolation of an Integrated Provirus of Moloney Murine Leukemia Virus with Long Terminal Repeats in Inverted Orientation: Integration Utilizing Two U3 Sequences

JOHN COLICELLI† AND STEPHEN P. GOFF*

Department of Biochemistry and Molecular Biophysics, Columbia University College of Physicians and Surgeons,
New York, New York 10032

Received 27 July 1987/Accepted 14 October 1987

We have previously described the construction of a mutant of Moloney murine leukemia virus, *in594-2*, which carries a 2-base-pair insertion in the U5 region of the genome and is partially defective in forming the integrated proviral DNA. We have now recovered a cloned copy of an unusual provirus from rat cells infected with this mutant. The viral genome is flanked by long terminal repeats in inverted orientation, with U3 sequences joined to cellular DNA at both of the outer edges. In addition, the provirus is a recombinant, containing a segment of a VL30 element in inverted orientation in place of the Moloney murine leukemia virus *env* region. The recovery of this provirus indicates that two U3 regions can be used for viral integration and suggests that there may be no absolute requirement in the reaction for those U5 sequences outside the 13-base-pair inverted repeats.

Soon after retroviral infection, a double-stranded DNA copy of the viral genome is synthesized and permanently inserted into the host cell DNA (see reference 19 for a review). The integration reaction occurs at specific sequences on the viral genome, such that the integrated provirus is always joined to the cell DNA in the same way. The ends of the viral genome consist of identical long terminal repeat (LTR) sequences in direct orientation; these LTRs are divided into three subregions in the order 5'-U3-R-U5-3' based on their contribution to viral genomic RNA transcripts. The organization of the integrated provirus means that at the 5' end of the genome, host sequences are joined to the U3 segment and that at the 3' end of the genome, host sequences are joined to the U5 segment. The outermost edges of the two LTRs (i.e., the 5' edge of U3 and the 3' edge of U5), and therefore the outer edges of the whole provirus, consist of perfect or nearly perfect short inverted repeats. The integration reaction thus apparently makes use of two similar sequences, in inverted orientation, at the outermost edges of the two LTRs.

Mutational studies in the avian (2, 7) and spleen necrosis (14) virus systems have shown that the terminal inverted repeat sequences, as well as additional flanking sequences, are required for the integration reaction. Deletion analysis of sequences near the inverted repeats of avian leukosis virus has shown that the required region is asymmetric: 18 to 22 base pairs (bp) of U5 and only 3 to 8 bp of U3 are required (2). Similar analysis of spleen necrosis virus has suggested that no more than 8 bp of U5 and 12 bp of U3 are necessary (14). In both systems, the essential region includes the inverted repeats as well as bases unique to U3 and U5. Preliminary work on the Moloney murine leukemia virus (M-MuLV) system suggests that few, if any, nucleotides outside the inverted repeats are required; the unique bases outside the unusually long inverted repeats (13 bp each) seem dispensable (J. E. Murphy and S. P. Goff, manuscript

in preparation). It is unclear in the murine system whether the unique portions of U5 and U3 are important in orienting the viral termini or whether any pair of inverted repeat sequences could be used.

To determine the extent of the sequences required for integrative recombination, we generated and analyzed a series of deletion, insertion, and substitution mutations at the U5 edge of M-MuLV (3; J. Colicelli and S. P. Goff, *J. Mol. Biol.*, in press). Large deletions in this region abolished integration, but substitutions and insertions in general had only small or no effects on the reaction. The analysis of the structure of the proviruses formed by some of these viable mutants has helped reveal some properties of the reaction (Colicelli and Goff, in press). We have now recovered an aberrant proviral clone of such an insertion mutant, whose structure suggests that two U3 sequences can be used in an integration reaction and that the unique sequences of U5 may not be essential.

Normal rat kidney cells were infected with a virus stock of mutant *in594-2*, bearing a 2-bp insertion at the U5 LTR edge, and after passage of the cells to permit proviral integration, genomic DNA was isolated from the cell population. A genomic library was made in the lambda phage vector Charon 4A after cleavage with *EcoRI*, and a single clone containing an apparently full-length provirus was isolated after screening 5×10^5 plaques with an M-MuLV-specific hybridization probe (plasmid pMS2; a gift of I. M. Verma). Restriction mapping of this clone, *in594-2IVL*, revealed that the LTRs were in inverted, rather than direct, orientation, with U3 sequences joined to cell DNA at both outer edges of the provirus (Fig. 1). In addition, numerous restriction site differences between the clone and wild-type M-MuLV were found in the viral *env* gene region. The provirus was bordered by short flanking sequences totaling only about 500 bp in length.

To understand the structure of this clone in greater detail, we determined the nucleotide sequences (12) of selected regions (A through D, Fig. 1). Both U5 edges, present at the inner edges of the LTRs, still retained the original 2-bp insertion mutation of the parent *in594-2* virus (regions A and

* Corresponding author.

† Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

FIG. 1. Structure of the cloned *in594-21VL* provirus. The inverted arrangement of the LTRs is indicated by the arrows and by the order of the U3, R, and U5 regions. The positions of the M-MuLV *gag* and *pol* genes and the inverted VL30 substitution are indicated by the boxes. The bars above the map, labeled A through F, indicate the positions of the sequence data in Fig. 2. Letters below the map indicate restriction enzyme cleavage sites as follows: R, *EcoRI*; P, *PstI*; X, *XhoI*; S, *Sall*; H, *HindIII*. kb, Kilobase.

B, Fig. 1 and 2). Both U5 edges abutted sequences that resembled primer-binding sites; the left U5 edge was adjacent to a complete binding site for the normal proline tRNA, whereas the right U5 edge was adjacent to an incomplete site. Both U3 edges, present at the outer edges of the LTRs, were joined to nonviral sequences. These edges had each lost their terminal 2 bp, typical of normal integration products, and were flanked by a 5-bp direct duplication, rather than the usual 4 bp (regions C and D, Fig. 1 and 2). Whereas the length of the target site duplication for M-MuLV is normally 4 bp, there are precedents for 5-bp duplications in proviruses of wild-type (18) and mutant (3, 9) MuLVs.

The sequence data suggested that a provirus with inverted LTRs had somehow formed and then integrated into rat DNA. To confirm that the apparent target site was present in uninfected rat cells and that the two sequences flanking the provirus were joined together before infection, we carried out Southern blot hybridizations (17). Subclones of the two flanking sequences were prepared, labeled by nick translation, and used to probe uninfected normal rat kidney cell DNA cleaved with four different enzymes. The results showed that the sequences found on opposite sides of the provirus were both single-copy sequences and were indeed linked on a single fragment in DNA of uninfected cells after cleavage with *XbaI*, *HindIII*, or *BamHI* (Fig. 3A). A single fragment of approximately 550 bp, the combined length of the two flanking sequences, was produced after cleavage with *EcoRI* (blot with 5' probe shown in Fig. 3B).

To determine the source of the new material with restriction site polymorphisms in the *env* region, we obtained sequence data at the boundaries of this area (regions E and F, Fig. 1 and 2). A short stretch of the sequence immediately flanking the right LTR and its partial primer-binding site proved to be a near-perfect (49 of 52) match to the sequence of a portion of the 5' end of the rat endogenous VL30 element, in inverted orientation (Fig. 2). VL30 elements are multicopy sequences found in the genomes of most mouse and rat strains (11); recombinants of VL30 and MuLV genomes have been detected previously (10). The Harvey and Kirsten transforming viruses are other examples of such recombinants, containing both MuLV and rat VL30 sequences (1, 20). The region of *in594-21VL* with homology to the VL30 sequence, reported in the context of the 5' end of the hybrid genome of the Harvey sarcoma virus (1, 6, 20), can be extended 10 bp further if a large gap is introduced, but the homology ends immediately thereafter. More extensive sequence data of the rest of the VL30 elements are not available. Sequence analysis at the other boundary of the new material shows an open reading frame, again in inverted orientation, extending up to a breakpoint within the M-MuLV *env* gene in the normal orientation (region F, Fig. 1

and 2). All of the sequence 5' to this point is identical to that of wild-type M-MuLV. The sequence of amino acids encoded by the new open reading frame is weakly homologous to retroviral *gag* sequences; comparison with M-MuLV, for

FIG. 2. Nucleotide sequences of selected regions of the *in594-21VL* provirus. Both strands of the virus are shown. The positions of sequences A to F on the provirus are as indicated in Fig. 1. Boxed areas indicate the viral LTRs; arrowheads on the boxes point in the direction of the arrows on the LTRs in Fig. 1. In A and B, primer-binding sites (pbs) for proline tRNA are indicated. In C and D, arrows mark the 5-bp direct duplication of target sequences at the outer edges of the provirus. In E, the sequence near the inverted 3' LTR (upper two lines) is compared with the VL30 sequences present in Harvey sarcoma virus (6) in inverted orientation. Matches between the two sequences are indicated by asterisks. In F, the recombination breakpoint is indicated by head-to-head arrows. The M-MuLV *env* sequence extends to nucleotide 6420, counted from the left edge of the left LTR. The putative VL30 sequences, in inverted orientation, are divided into codons of the open reading frame, and the predicted amino acid sequence of the protein encoded by this region is shown underneath. Amino acids 6 to 26 of this sequence, from the N terminus to the C terminus, align weakly with the P30^{KK} protein of M-MuLV (beginning at residue 218) and simian sarcoma virus (beginning at residue 225).

FIG. 3. Southern blots of uninfected rat DNA. Subclones of the 5'-flanking region (*Eco*RI to *Nhe*I) and the 3'-flanking region (*Nhe*I to *Eco*RI) of clone *in594-2IVL* were prepared in plasmid vector pSP64, and the excised fragments, 230 and 250 bp in length, were labeled by nick translation for use as hybridization probes. (A) Agarose gel electrophoresis was performed after digestion of normal rat kidney cell DNA with *Xba*I (lanes X), *Hind*III (lanes H), or *Bam*HI (lanes B), and duplicate blots were hybridized with the two flanking probes. The positions of migration of size standards are indicated. (B) Agarose gel electrophoresis was performed after normal rat kidney cell DNA was digested with *Eco*RI, and blots were hybridized as before. Lane R shows the result of probing with the 5'-flanking probe. The position of the single 550-bp band is indicated. kb, Kilobases.

example, shows an 11 of 29 match (with a 1-residue gap) with a sequence near the C terminus of P30^{uag} (16). The sequence is a slightly closer match (14 of 29 with no gaps) to the same region of the simian sarcoma virus *gag* gene (5). Although VL30 *gag* sequences have not been reported, the position of the homology to known *gag* sequences suggests that the entire 1.8 kilobases of new material could consist of a single large patch of VL30-related sequence, extending roughly from the edge of the LTR through P30^{uag}.

How can we account for the appearance of this provirus? We need to explain how the two inverted M-MuLV LTRs arose and also how the VL30 sequences were acquired; an intriguing aspect of the latter sequences is that one endpoint lies within the primer-binding site of the 3' LTR. A simple model of the events that might have generated the structure is shown in Fig. 4. We suppose that, during the spread of mutant *in594-2* in the infected population, there was pseudotype formation that resulted in frequent transfer of VL30 RNAs from cell to cell (15). A cell infected with two virions, one containing a heterodimeric viral RNA (including *in594-2* and VL30 RNAs) and one containing only mutant *in594-2*, could have formed the provirus. First, we propose that in the mixed virion, minus-strand synthesis proceeded normally on both the VL30 and *in594-2* genomes (Fig. 4, line 2), giving as usual a nearly full-length DNA with the primer tRNA at the

5' end and a sequence complementary to the primer-binding site at the 3' end (8). Plus-strand synthesis would give the usual plus-strand strong stops, copying part of the tRNA. Then, in the heterodimer reverse transcription complex, we propose that a plus-strand strong stop from the *in594-2* genome was able to jump or translocate in *trans* to the VL30 minus strand; completion of synthesis after such a jump would yield a recombinant genome with an M-MuLV LTR at the 5' end, still containing the *in594-2* mutation, on a VL30 body (Fig. 4, line 3, left). The recombination point here would lie somewhere within the primer-binding site, normally the sequence used to mediate this jump (8). Such a *trans* jump in mutant *in594-2* might be promoted by a defect in removal of the tRNA primer, a step normally required for that jump; indeed, this mutant has yielded aberrant DNAs that retain tRNA sequences after its reverse transcription (4). We should note that recombination endpoints at the LTR edge, rather than in the primer-binding site, have been reported and explained by a different mechanism (13). We next propose that in parallel, in the other virion present in the infected cell, reverse transcription of the *in594-2* genome occurred normally (Fig. 4, line 3, right). The second recombination endpoint in the hybrid provirus would be created by a nonhomologous illegitimate crossover: the two linear viral genomes would recombine in inverted orientation at the

FIG. 4. Model for the formation of integrant *in594-2IVL*. VL30 sequences are indicated by wavy lines, and M-MuLV sequences are indicated by straight lines. The position of the 2-bp insertion mutation is indicated by a black triangle. See the text for a detailed explanation.

position of the two X's to form a hybrid genome (Fig. 4, line 4). At the outer edges of this linear genome would be intact wild-type U3 regions, and at the inner edges would be mutant U5 regions bearing the *in594-2* mutation. Finally, the integration of this provirus, either directly or after cyclization, would yield the structure recovered in clone *in594-2IVL*.

The structure of the provirus in clone *in594-2IVL* strongly suggests that two U3 edges can cooperate to form an acceptable substrate for integrative recombination. Thus, it is apparently not essential for the reaction that U5 sequences be present, at least not in the vicinity of the breakpoints. Is the DNA with inverted LTRs an efficient substrate? The normal proviral structure of mutant *in594-2* can be integrated, and the virus can give rise to progeny after infection with no significant delay compared with the wild type (Colicelli and Goff, in press). Nevertheless, many of the DNAs formed by this mutant during its reverse transcription are aberrant, carrying extra bases at the 3' end of the genome (4; Colicelli and Goff, in press), and some of these DNAs may integrate poorly. We propose that the DNA with inverted LTRs was detected because of its selective advantage: that it is a better substrate for integration than the DNA with a mutant U5 edge. We imagine that generation of the aberrant precursor as outlined in Fig. 4 is rare, and so the recovery of even one clone of such an integrant, with no particular selection or screen, implies that it might be a good substrate. Finally, we should note that the observed structure of the provirus suggests that it is defective and could not be transmitted further; even if the genome were transcribed and packaged into virions, it could not be reverse transcribed in a subsequent infection.

This work was supported by Public Health Service award CA 30488 from the National Cancer Institute.

LITERATURE CITED

- Chien, Y.-H., M. Lai, T. Y. Shih, I. M. Verma, E. M. Scolnick, P. Roy-Burman, and N. Davidson. 1979. Heteroduplex analysis of the sequence relationships between the genomes of Kirsten and Harvey sarcoma viruses, their respective parental murine leukemia viruses, and the rat endogenous 30S RNA. *J. Virol.* **31**:752-760.
- Cobrinik, D., R. Katz, R. Terry, A. M. Skalka, and J. Leis. 1987. Avian sarcoma and leukemia virus *pol*-endonuclease recognition of the tandem long terminal repeat junction: minimum site required for cleavage is also required for viral growth. *J. Virol.* **61**:1999-2008.
- Colicelli, J., and S. P. Goff. 1985. Mutants and pseudorevertants of Moloney murine leukemia virus with alterations at the integration site. *Cell* **42**:573-580.
- Colicelli, J., and S. P. Goff. 1986. Structure of a cloned circular retroviral DNA containing a tRNA sequence between the terminal repeats. *J. Virol.* **57**:674-677.
- Devare, S. G., E. P. Reddy, J. D. Law, K. C. Robbins, and S. A. Aaronson. 1983. Nucleotide sequence of the simian sarcoma virus genome: demonstration that its acquired cellular sequences encode the transforming gene product p28^{sis}. *Proc. Natl. Acad. Sci. USA* **80**:731-735.
- Dhar, R., R. W. Ellis, T. Y. Shih, S. Oroszlan, B. Shapiro, J. Maizel, D. Lowy, and E. Scolnick. 1982. Nucleotide sequence of the oncogene encoding the p21 transforming protein of Harvey murine sarcoma virus. *Science* **217**:934-937.
- Duyk, G., M. Longiaru, D. Cobrinik, R. Kowal, P. deHaseth, A. M. Skalka, and J. Leis. 1985. Circles with two tandem long terminal repeats are specifically cleaved by *pol* gene-associated endonuclease from avian sarcoma and leukemia viruses: nucleotide sequences required for site-specific cleavage. *J. Virol.* **56**:589-599.
- Gilboa, E., S. W. Mitra, S. Goff, and D. Baltimore. 1979. A detailed model of reverse transcription and tests of crucial aspects. *Cell* **18**:93-100.
- Hagino-Yamagishi, K., L. A. Donehower, and H. E. Varmus. 1987. Retroviral DNA integrated during infection by an integration-deficient mutant of murine leukemia virus is oligomeric. *J. Virol.* **61**:1964-1971.
- Itin, A., and E. Keshet. 1983. Apparent recombinants between virus-like (VL30) and murine leukemia virus-related sequences in mouse DNA. *J. Virol.* **47**:178-184.
- Keshet, E., Y. Shaul, J. Kaminchik, and H. Aviv. 1980. Heterogeneity of virus-like genes encoding retrovirus-associated 30S RNA and their organization within the mouse genome. *Cell* **20**:431-439.
- Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499-560.
- Olsen, J. C., and R. Swanstrom. 1985. A new pathway in the generation of defective retrovirus DNA. *J. Virol.* **56**:779-789.
- Panganiban, A. T., and H. M. Temin. 1983. The terminal nucleotides of retrovirus DNA are required for integration but not virus production. *Nature (London)* **306**:155-160.
- Scolnick, E. M., W. C. Vass, R. S. Howk, and P. H. Duesberg. 1979. Defective retrovirus-like 30S RNA species of rat and mouse cells are infectious if packaged by type C helper virus. *J. Virol.* **29**:964-972.
- Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe. 1981. Nucleotide sequence of Moloney murine leukemia virus. *Nature (London)* **293**:543-548.
- Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503-517.
- Van Beveren, C., E. Rands, S. K. Chattopadhyay, D. R. Lowy, and I. M. Verma. 1981. Long terminal repeat of murine retroviral DNAs: sequence analysis, host-proviral junctions, and preintegration site. *J. Virol.* **41**:542-556.
- Varmus, H. E. 1983. Retroviruses, p. 411-503. In J. A. Shapiro (ed.), *Mobile genetic elements*. Academic Press, Inc., New York.
- Young, H. A., M. A. Gonda, D. DeFeo, R. W. Ellis, K. Nagashima, and E. Scolnick. 1980. Heteroduplex analysis of cloned rat endogenous replication-defective (30S) retrovirus and Harvey murine sarcoma virus. *Virology* **107**:89-99.