Plus-Strand Origin for Human Immunodeficiency Virus Type 1: Implications for Integration

KATHERINE A. PULLEN AND JAMES J. CHAMPOUX*

Department of Microbiology, School of Medicine, University of Washington, Seattle, Washington 98195

Received 5 June 1990/Accepted 23 August 1990

The start site for human immunodeficiency virus type 1 plus strands within the polypurine tract was mapped by an in vitro analysis to the sequence 5'-ACTG From this result, it can be inferred that integration of human immunodeficiency virus type 1 must be accompanied by the loss of two base pairs from the polypurine tract-primed long terminal repeat end.

Retroviral RNA replication occurs by way of a DNA intermediate that becomes stably integrated into the host cell DNA (36). The DNA intermediate produced by reverse transcription is bounded by long terminal repeats (LTRs) that contain the sequence repeated at the ends of the viral RNA (R) as well as sequences unique to the 5' end (U5) and the 3' end (U3) of the viral genome (LTR = U3-R-U5, Fig. 1). During reverse transcription, each DNA strand is initiated at a unique locus determined by the 3' end of an RNA primer. The minus DNA strand that is complementary to the viral RNA is primed by a tRNA molecule that is paired with the RNA at a site near its 5' terminus called the primer-binding site. Plus strands are primed by a specific RNase H fragment derived from a region of the viral RNA near its 3' terminus called the polypurine tract (PPT) (7, 16, 17, 23-25, 33, 35). Based on the prevailing model for reverse transcription (12), the left and right ends of the LTRs are determined by the sites of initiation of the plus and minus DNA strands, respectively.

Integration requires the viral IN protein and the short inverted repeat sequences (or imperfect inverted repeats) found at the termini of the linear product of reverse transcription (2-6, 9, 10, 13, 18, 19, 21, 26, 29). Prior to the characterization of the lentiviruses and the more complex oncogenic retroviruses (e.g., human T-cell lymphotropic virus type I), all the known inverted repeats found at the LTR termini began with the sequence 5'-AATG Since all integrated proviruses begin with a TG dinucleotide and end with a CA dinucleotide, it appeared that two A · T base pairs were symmetrically lost from each end of the linear DNA during the integration process (36). Consistent with this observation, recent studies on the mechanism of Moloney murine leukemia virus (M-MuLV) integration have revealed that the initial step in the integration reaction is the removal of two 3' nucleotides (thymidine residues) from each end of the linear DNA (3, 10, 15, 26).

Examination of the sequence adjacent to the human immunodeficiency virus type 1 (HIV-1) primer-binding site (22) initially suggested that the model described above for the initial steps of integration does not hold for all retroviruses (Fig. 1). For HIV-1, the tRNA-primed minus strand is predicted to begin with the sequence 5'-CTG . . . and therefore one C · G base pair rather than two A · T base pairs should be removed from the right LTR end during integration. From an analysis of the circle junction formed by the

blunt-end ligation of the linear DNA, it has been possible to deduce the actual sequence present at the right LTR end (37; J. Kulkosky, R. A. Katz, and A. M. Skalka, J. Acquired Immun. Defic. Synd., in press). Although the predicted sequence (5'-CTG . . .) was occasionally found, the preponderant product had the sequence 5'-ACTG . . . (37; Kulkosky et al., in press). The most likely explanation for the discrepancy between the observed and predicted structures is that primer removal is incomplete, resulting in an extra ribonucleotide A residue on the 5' end of newly synthesized minus strands (37). Assuming the resulting structure is a substrate for integration, two base pairs rather than the predicted one base pair would have to be removed from the right end during integration.

By comparing the HIV-1 PPT sequence with the PPT sequences from other retroviruses, it is not possible to discern the precise start site for plus strands. The two most likely start sites are indicated by the arrows above the PPT in Fig. 1. Initiation between the A and C residues would generate plus strands that begin with the same sequence as predicted from the tRNA structure for minus strands (5'-CTG . . .) and the left LTR end would have the structure labeled I in Fig. 1. Alternatively, by analogy with M-MuLV and Rous sarcoma virus, initiation between the G and A residues would yield plus strands that begin with the sequence 5'-ACTG . . . and the left LTR end would have the structure labeled II in Fig. 1. The objective of the work described here was to distinguish between these two possibilities by mapping the HIV-1 plus-strand origin.

We have previously used purified components to reconstruct M-MuLV plus-strand priming in vitro (7, 23, 24). When presented with a DNA-RNA hybrid containing the PPT and all four deoxynucleoside triphosphates, the purified M-MuLV reverse transcriptase will initiate plus-strand synthesis at the correct site near the 3' end of the PPT. It appears that the RNase H activity of the reverse transcriptase is responsible for the endonucleolytic cleavage that creates the primer RNA. We used the same methodology to map the HIV-1 plus-strand origin. The 224-base-pair *Xho*I (8896)-to-*Eco*RV (9119) fragment from HIV-1 (Fig. 1) was cloned in M13mp7 in the orientation such that the packaged single-stranded DNA contained the viral minus strand. The insert was excised by cutting the single-stranded DNA in the paired polycloning region of the vector with *Eco*RI (1) and purified by sedimentation in alkaline sucrose. The *Xho*I (8896)-to-*Sac*I (9571) restriction fragment was cloned in pBSM13(+) and, after cutting with *Eco*RV, used as a template for T7 RNA polymerase to produce a runoff plus-strand

* Corresponding author.

FIG. 1. Replication and integration landmarks on the retroviral genome. The tRNA primer-binding site (PBS) and the PPT are located adjacent to the left and right LTR sequences, respectively, as shown in the top line. The relative locations of the *Xho*I site at position 8896 (X), the *Eco*RV site at 9119 (E), the sequencing oligonucleotide (9101 to 9119), and the PPT (9068 to 9083) are indicated. (Nucleotide numbering refers to the HIV-1 HXB2 sequence). The sequence of the LTR terminus that is predicted from the structure of the primer tRNA is shown below on the left side and is labeled right LTR end (see text for alternative possibilities). The two arrows above the PPT RNA sequence identify the two most likely sites for initiation of plus strands. Initiation at the second arrow (between the A and the C) would generate the left LTR end labeled I, whereas initiation at the first arrow (between the G and the A) would generate the left LTR end labeled II. The base pairs shown in boxes on the lower portion of the figure are the ones that must be removed during integration to generate the proper proviral ends.

RNA that is complementary to the HIV-1 sequences present in the minus-strand DNA fragment. The RNA and DNA fragments were hybridized and used as templates for DNA synthesis by HIV-1 reverse transcriptase (kindly provided by Lawrence Loeb, University of Washington) as previously described (7). The DNA products were treated with alkali (0.3 M NaOH, 65°C, 25 min) to remove any residual RNA primers, and the 5' ends were mapped by an oligonucleotide extension assay, using the DNA polymerase activity of M-MuLV reverse transcriptase (23). By comparing the mobility of the extension product (Fig. 2A, lane 1) with the adjacent sequencing ladders, it can be seen that HIV-1 plus strands start with the sequence 5'-ACTG... (structure II, bottom of Fig. 1). No bands were detectable in a control experiment from which the reverse transcriptase had been omitted (Fig. 2A, lane 2). The recently reported sequence data for HIV-1 circle junctions (37; Kulkosky et al., in press) corroborate our result concerning the site of initiation of plus strands.

As a positive control for the above experiment, the HIV-1 DNA-RNA hybrids were used as substrates for the M-MuLV reverse transcriptase (Fig. 2B). The HIV-1 PPT differs from the M-MuLV PPT at only one position; the residue 11 bases upstream of the start site is a G rather than an A. Because of our previous experiments in which we systematically studied the effects of single- and double-base changes on the specificity of M-MuLV plus-strand priming (24), we did not anticipate that this change would affect the

breakage specificity of the RNase H. The results indicate that the HIV-1 PPT is indeed properly recognized by the M-MuLV reverse transcriptase in the plus-strand priming reaction (Fig. 2B, lane 1). Since the sequences flanking the HIV-1 PPT are completely different from those flanking the M-MuLV PPT, this result further suggests that all the determinants of correct plus-strand priming reside within the PPT itself.

The location of the start site for HIV-1 plus strands has important implications for the process of integration. From the time of the initial studies on the oncovirus subgroup of retroviruses, it has generally been assumed that integration is accompanied by the symmetrical loss of two A · T base pairs from each end of the viral DNA. The results reported here for HIV-1 indicate that two base pairs are indeed removed from the left LTR end, but one is an A · T base pair and the other is a C · G base pair. The number of base pairs lost from the tRNA-primed right LTR end is likely also to be two, given the sequence data for the cloned circle junctions (37; Kulkosky et al., in press). Therefore, like the oncoviruses, the processing of the HIV-1 LTR termini in preparation for integration probably involves the symmetrical loss of the same two base pairs from each end.

Figure 3 shows a compilation of the sequences found at the left and right LTR ends for a variety of retroviruses. The sequences are aligned on the conserved TG and CA dinucleotides found at the termini of the integrated proviruses to facilitate the identification of those residues lost during

FIG. 2. Determination of the HIV plus-strand initiation site by oligonucleotide primer extension analysis. (A) A 224-base-pair DNA-RNA hybrid containing the HIV PPT was incubated with the HIV reverse transcriptase in the presence of all four deoxynucleoside triphosphates. The labeled minus-strand sequencing oligonucleotide (sequence 5'-ATCTTGTCTTCTTTGGGAG-3'; see Fig. 1 for location) was annealed to the alkali-treated products and extended by using the DNA polymerase activity of M-MuLV reverse transcriptase (Bethesda Research Laboratories) as described previously (23). The 5' ends of the plus strands were mapped by subjecting the primer extension products to electrophoresis in an 8% polyacrylamide-8 M urea gel alongside the products from the dideoxy sequencing reactions (lane 1). A parallel control sample from which the HIV reverse transcriptase had been omitted was treated identically (lane 2). (B) The same analysis was done with the HIV PPT DNA-RNA hybrids as templates for the M-MuLV reverse transcriptase instead of HIV reverse transcriptase (lane 1). The reverse transcriptase was omitted from the control reaction (lane 2). For both panels, the dideoxy sequencing ladders (T, C, A, G) were generated with Sequenase (U.S. Biochemicals Corp.) under the conditions recommended by the supplier, and the same labeled oligonucleotide annealed to a single-stranded M13mp7 template containing a plus-strand insert of the *Xho*I (8896)-to-*Eco*RV (9119) restriction fragment. The sequence of bases in the immediate vicinity of the plus-strand initiation site is given along the right side of each panel.

integration. For those viruses listed in Fig. 3A, the origins for plus-strand synthesis were determined either directly, by mapping the plus-strand start, or, for spleen necrosis virus, by inference from the sequence of the LTR circle junction (20). For those viruses listed in Fig. 3B, the plus-strand origins were not experimentally determined and therefore the placement of two extra base pairs at the left LTR end must be considered provisional. From an examination of the sequences in Fig. 3B, it appears that neither the nature nor the number of bases removed during integration need be the same on the two ends of the DNA. For visna virus, human T-cell lymphotropic virus type I, and bovine leukemia virus, two base pairs are probably removed from each end, but the specific bases are different. For HIV-2 and simian immunodeficiency virus (MM251), it is likely that processing is asymmetrical, with three bases being removed from the right end and two being removed from the left end. Finally, simian immunodeficiency virus (TYO) appears to be similar to HIV-1, with the symmetry of the reaction depending on whether or not an extra ribonucleotide is left on the 5' end of the nascent minus strands during removal of the tRNA primer. The possibility that different base pairs and different numbers of base pairs can be present on the two ends of the integrating retroviral DNA suggests that the integration

VIRUS	LEFT LTR END		RIGHT LTR END	
	U3		U5	
A				
M-MuLV	AA	TGAAAGACCCC.....	GGGGTCTTTCA	TT
	TT	ACTTTCTGGG.....	CCCCAGAAAGT	AA
RSV	AA	TGTAGTCTTAT.....	AGAAGGCTTCA	TT
	TT	ACATCAGAATA.....	TCTTCGGAAGT	AA
SNV	AA	TGTGGGAGGGA.....	TCGGTACAACA	TT
	TT	ACACCTCCCT.....	AGCCATGTTGT	AA
HIV-1	AC	TGGAAGGGCTA.....	AATCTCTAGCA	G
	TG	ACCTTCCCGAT.....	TTAGAGATCGT	C
B				
EIAV	?	AC TGTGGGGTTTT.....	GAGATCCTACA	GT
	TG	ACACCCCAAAA.....	CTCTAGGATGT	CA
FIV	?	AC TGGGATGAGTA.....	AGAAGTTCGCA	GT
	TG	ACCCTACTCAT.....	TCTTGAAGCGT	CA
VISNA	?	AC TGTGAGGACAG.....	CGGATCTCGCA	GC
	TG	ACAGTCCTGTC.....	GCCTAGAGCGT	CG
HTLV-1	?	GA TGACAATGACC.....	TTTGTACACA	GT
	CT	ACTGTTACTGG.....	AAATCATGTGT	CA
BLV	?	TT TGTATGAAAGA.....	ACCGGCAACA	AT
	AA	ACATACTTTCT.....	TGGCCGTTTGT	TA
HIV-2	?	AC TGGAAGGGCTG.....	AATCCCTAGCA	GGT
	TG	ACCTTCCCGAC.....	TTAGGGATCGT	CCA
SIV(MM251)	?	AC TGGAAGGGATT.....	AATCCCTAGCA	GAT
	TG	ACCTTCCCTAA.....	TTAGGGATCGT	CTA
SIV(TYO)	?	AC TGGATGGGATT.....	AAATCCAGCA	G
	TG	ACCTACCTTAA.....	TTTGAAGTCTG	C

FIG. 3. Retroviral LTR termini. The sequences found at the ends of the linear DNAs produced by reverse transcription for a variety of retroviruses are compared. The sequences are aligned on the 5'-TG...CA-3' residues that form the ends of the integrated proviral DNAs. Thus, the base pairs outside the two vertical lines are the ones removed during integration. In all cases, the right LTR end (U5) was inferred from the sequences in the RNA genome that are located immediately downstream from the point where the CCA end (3' end) of the tRNA primer is paired. With HIV-1, this approach may lead to an erroneous conclusion concerning the right LTR end sequence (see text). For those viruses listed in panel A, the site of initiation of plus strands was determined, allowing an unambiguous determination of the left LTR end (U3). For the viruses shown in panel B, the exact start site for plus-strand synthesis was not determined directly, but in accord with the known cases, the sequences are written with two base pairs preceding the TG dinucleotide. The question marks serve to emphasize the provisional nature of this assignment. Abbreviations and references: M-MuLV, Moloney murine leukemia virus (31); RSV, Rous sarcoma virus (28); SNV, spleen necrosis virus (20); HIV-1, human immunodeficiency virus type 1 strain HXB2 (22); EIAV, equine infectious anemia virus (14); FIV, feline immunodeficiency virus (34); VISNA, visna lentivirus (32); HTLV-I, human T-cell lymphotropic virus type I (30); BLV, bovine leukemia virus (27); HIV-2, human immunodeficiency virus type 2, isolate Hiv2fg (38); SIV(MM251), simian immunodeficiency virus, isolate MM251 (8); SIV(TYO), simian immunodeficiency virus, isolate TYO-1 (11).

machinery does not interact in a sequence-specific fashion with the very ends of the LTRs. A similar conclusion has been reached for M-MuLV on the basis of the finding that mutants with quantitative and qualitative alterations at the right LTR end can still integrate in vivo. (4, 5).

This research was supported by Public Health Service grant CA 51605 from the National Cancer Institute.

We thank Lance Ishimoto, Paul Krogstad, and Knut Madden for technical advice and for helpful comments during the preparation of the manuscript. We are grateful to Lawrence Loeb for generously providing purified HIV-1 reverse transcriptase.

LITERATURE CITED

1. Been, M. D., and J. J. Champoux. 1983. Cutting M13mp7 phage DNA and excision of cloned single-stranded sequences by restriction endonucleases. *Methods Enzymol.* **101**:90-98.
2. Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop. 1987. Correct integration of retroviral DNA in vitro. *Cell* **49**:347-356.
3. Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop. 1989. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc. Natl. Acad. Sci. USA* **86**:2525-2529.
4. 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.
5. Colicelli, J., and S. P. Goff. 1988. Sequence and spacing requirements of retrovirus integration site. *J. Mol. Biol.* **199**:47-59.
6. Donehower, L. A., and H. E. Varmus. 1984. A mutant murine leukemia virus with a single missense codon in *pol* is defective in a function affecting integration. *Proc. Natl. Acad. Sci. USA* **81**:6461-6465.
7. Finston, W. I., and J. J. Champoux. 1984. RNA-primed initiation of Moloney murine leukemia virus plus strands by reverse transcriptase in vitro. *J. Virol.* **51**:26-33.
8. Franchini, G., C. Gurgo, H.-G. Guo, R. C. Gallo, E. Collati, K. A. Fargnoli, L. F. Hall, F. Wong-Staal, and M. S. Reitz, Jr. 1987. Sequence of simian immunodeficiency virus and its relationship to the human immunodeficiency viruses. *Nature (London)* **328**:539-543.
9. Fujiwara, T., and R. Craigie. 1989. Integration of mini-retroviral DNA: a cell-free reaction for biochemical analysis of retroviral integration. *Proc. Natl. Acad. Sci. USA* **86**:3065-3069.
10. Fujiwara, T., and K. Mizuuchi. 1988. Retroviral DNA integration: structure of an integration intermediate. *Cell* **54**:497-504.
11. Fukasawa, M., T. Miura, A. Hasegawa, S. Morikawa, H. Tsujimoto, K. Miki, T. Kitamura, and M. Hayami. 1988. Sequence of simian immunodeficiency virus from African green monkey, a new member of HIV/SIV group. *Nature (London)* **333**:457-461.
12. Gilboa, E., S. W. Mitra, S. Goff, and D. Baltimore. 1980. A detailed model of reverse transcription and test of crucial aspects. *Cell* **18**:93-100.
13. Katzman, M., R. A. Katz, A. M. Skalka, and J. Leis. 1989. The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the in vivo sites of integration. *J. Virol.* **63**:5319-5327.
14. Kawakami, T., L. Sherman, J. Dahlberg, A. Gazit, A. Yaniv, S. R. Tronick, and S. A. Aaronson. 1987. Nucleotide sequence analysis of equine infectious anemia virus proviral DNA. *Virology* **158**:300-312.
15. Loebel, L. I., J. E. Murphy, and S. P. Goff. 1989. The palindromic LTR-LTR junction of Moloney murine leukemia virus is not an efficient substrate for proviral integration. *J. Virol.* **63**:2629-2637.
16. Mitra, S. W., M. Chow, J. Champoux, and D. Baltimore. 1982. Synthesis of murine leukemia virus plus strong stop DNA initiates at a unique site. *J. Biol. Chem.* **257**:5983-5986.
17. Omer, C. A., R. Resnick, and A. J. Faras. 1984. Evidence for involvement of an RNA primer in initiation of strong-stop plus DNA synthesis during reverse transcription in vitro. *J. Virol.* **50**:465-470.
18. 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.
19. Panganiban, A. T., and H. M. Temin. 1984. The retrovirus *pol* gene encodes a product required for DNA integration: identification of a retrovirus *int* locus. *Proc. Natl. Acad. Sci. USA* **81**:7885-7889.
20. Panganiban, A. T., and H. M. Temin. 1984. Circles with two tandem LTRs are precursors to integrated retrovirus DNA. *Cell* **36**:673-679.
21. Quinn, T. P., and D. P. Grandgenett. 1988. Genetic evidence that the avian retrovirus DNA endonuclease domain of *pol* is necessary for viral integration. *J. Virol.* **62**:2307-2312.
22. Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J. A. Rafalski, E. A. Whitehorn, K. Baumeister, L. Ivanoff, S. R. Petteway, Jr., M. L. Pearson, J. A. Lautenberger, T. S. Papas, J. Ghrayeb, N. T. Chang, R. C. Gallo, and F. Wong-Staal. 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature (London)* **313**:277-284.
23. Rattray, A. J., and J. J. Champoux. 1987. The role of Moloney murine leukemia virus RNase H activity in the formation of plus-strand primers. *J. Virol.* **61**:2843-2851.
24. Rattray, A. J., and J. J. Champoux. 1989. Plus-strand priming by Moloney murine leukemia virus: the sequence features important for cleavage by RNase H. *J. Mol. Biol.* **208**:445-456.
25. Resnick, R., C. A. Omer, and A. J. Faras. 1984. Involvement of retrovirus reverse transcriptase-associated RNase H in the initiation of strong-stop (+) DNA synthesis and the generation of the long terminal repeat. *J. Virol.* **51**:813-821.
26. Roth, M. J., P. L. Schwartzberg, and S. P. Goff. 1989. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell* **58**:47-54.
27. Sagata, N., T. Yasunaga, J. Tsuzuku-Kawamura, K. Ohishi, Y. Ogawa, and Y. Ikawa. 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. *Proc. Natl. Acad. Sci. USA* **82**:677-681.
28. Schwartz, D. E., R. Tizard, and W. Gilbert. 1983. Nucleotide sequence of Rous sarcoma virus. *Cell* **32**:853-869.
29. Schwartzberg, P., J. Colicelli, and S. P. Goff. 1984. Construction and analysis of deletion mutations in the *pol* gene of Moloney murine leukemia virus: a new viral function required for productive infection. *Cell* **37**:1043-1052.
30. Seiki, M., S. Hattori, Y. Hirayama, and M. Yoshida. 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc. Natl. Acad. Sci. USA* **80**:3618-3622.
31. Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe. 1981. Nucleotide sequence of Moloney murine leukemia virus. *Nature (London)* **293**:543-548.
32. Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P. Tiollais, A. Haase, and S. Wain-Hobson. 1985. Nucleotide sequence of visna lentivirus: relationship to the AIDS virus. *Cell* **42**:369-382.
33. Sorge, J., and S. H. Hughes. 1982. Polypurine tract adjacent to the U3 region of the Rous sarcoma virus genome provides a *cis*-acting function. *J. Virol.* **43**:482-488.
34. Talbot, R. L., E. E. Sparger, K. M. Lovelace, W. M. Fitch, N. C. Pedersen, P. A. Luciw, and J. H. Elder. 1989. Nucleotide sequence and genomic organization of feline immunodeficiency virus. *Proc. Natl. Acad. Sci. USA* **86**:5743-5747.
35. Taylor, J., and L. Sharmeen. 1987. Retrovirus genome replication: priming specificities of plus-strand DNA synthesis. *J. Cell Sci. Suppl.* **7**:189-195.
36. Varmus, H., and P. Brown. 1989. Retroviruses, p. 53-108. In D. E. Berg and M. M. Howe (ed.), *Mobile DNA*. American Society for Microbiology, Washington, D.C.
37. Whitcomb, J. M., R. Kumar, and S. H. Hughes. 1990. Sequence of the circle junction of human immunodeficiency virus type 1: implications for integration. *J. Virol.* **64**:4903-4906.
38. Zagury, J. F., G. Franchini, M. Reitz, E. Collalti, B. Starcich, L. Hall, K. Fargnoli, L. Jagodzinski, H. G. Guo, and F. Laure. 1988. Genetic variability between isolates of human immunodeficiency virus (HIV) type 2 is comparable to the variability among HIV type 1. *Proc. Natl. Acad. Sci. USA* **85**:5941-5945.