The Sequence of Human Immunodeficiency Virus Type 2 Circle Junction Suggests That Integration Protein Cleaves the Ends of Linear DNA Asymmetrically

JEANNETTE M. WHITCOMB AND STEPHEN H. HUGHES*

*ABL-Basic Research Program, National Cancer Institute-Frederick Cancer Research and Development Center,
P.O. Box B, Frederick, Maryland 21702-1201*

Received 28 January 1991/Accepted 2 April 1991

The sequence of the human immunodeficiency virus type 2 circle junction was determined. The most common sequence found between the conserved CA and TG dinucleotides at the ends of the integrated provirus was five bases long (GGTAC). This suggests that the integration of human immunodeficiency virus type 2 DNA is accompanied by the asymmetric loss of two and three bases, respectively, from the U3 and U5 ends of the linear double-stranded DNA prior to integration.

A key step in the early stage of the retroviral life cycle is the reverse transcription of the viral RNA genome into a linear double-stranded DNA (dsDNA) molecule and the integration of this linear dsDNA into the host genome (30, 33). These two events, unique to retroviruses and retrotransposons, should be excellent targets for interrupting the retroviral life cycle. Understanding the molecular basis of these steps in the replication of human immunodeficiency virus (HIV) may be of great importance in the design of successful therapies against AIDS.

The generation of the linear dsDNA molecule by reverse transcription involves a complex series of events (30, 33). The structures of the viral RNA, the linear dsDNA, and the integrated proviral DNA are shown diagrammatically in Fig. 1. The reverse transcription reaction generates terminally redundant sequences known as long terminal repeats (LTRs). The initiation of plus- and minus-strand DNA synthesis at discrete sites within the RNA genome defines the boundaries of the LTRs (Fig. 1). Minus-strand synthesis is initiated from a cellular tRNA primer that is annealed to a specific complementary primer-binding site (PBS) located immediately adjacent to the U5 sequence. Plus-strand DNA synthesis is initiated from an RNA primer that is generated by a specific RNase H cleavage of the RNA genome within a polypurine tract (PPT) adjacent to U3. The most widely accepted model of reverse transcription dictates that the sites of initiation of both plus- and minus-strand synthesis precisely define the ends of the unintegrated linear dsDNA molecule. More specifically, the first DNA base added to the tRNA primer defines the end of the U5 region in the right LTR and the subsequent removal of this tRNA primer by RNase H exactly defines the right end of the linear dsDNA molecule. The left end of the linear dsDNA molecule is defined by two RNase H events; the first generates the RNA primer that initiates plus-strand synthesis, and the second removes the RNA primer from the left end of the DNA molecule.

Following reverse transcription, the linear dsDNA molecule, which is bounded by LTRs, is transported to the nucleus, where it is the precursor to the integrated provirus.

The linear molecule also gives rise in the nucleus to two circular forms of DNA that contain either one or two LTRs.

The basic mechanism of retroviral integration is emerging from in vivo and in vitro studies of several retroviruses (1-4, 7-12, 16, 26). It is now clear that the linear viral DNA molecule, and not one of the circular forms, is the direct precursor to the integrated provirus (1, 2, 8, 19). A virally encoded product, integration protein (IN), recognizes sequences at the ends of the linear molecule and removes a specific number of bases, usually two, from each of the 3' ends of the linear viral DNA (Fig. 1) (7, 10, 16, 17, 31). The IN then joins the recessed viral 3' ends to the 5' ends of a double-stranded staggered cut in the host DNA. This intermediate is repaired, presumably by host enzymes. The resulting integrated provirus is shorter than the linear dsDNA molecule by the number of bases removed by IN prior to integration (Fig. 1).

Studies with various mutant forms of retroviruses have shown that the terminal 15 to 20 bases at each end of the linear DNA are important for correct integration to occur (3, 6, 17, 21, 26). However, the specificity of the cleavage event by IN is not completely understood. To fully understand the integration reaction, it is important to define the exact substrate for IN (i.e., the ends of the linear dsDNA molecule).

Figure 2 compares the sequences of the genome/LTR boundaries of a number of retroviruses. Examination of these sequences should allow us to predict the sequences present at the ends of the linear substrate for integration. For example, the right end of the linear dsDNA molecule should contain the U5 sequences adjacent to the PBS (Fig. 2, box). It should also be possible to predict the bases removed by IN prior to integration. All known integrated proviruses are bounded on the left (U3) end by a TG dinucleotide and on the right (U5) end by a CA dinucleotide (Fig. 2, shaded boxes). The bases present between the tRNA primer (boxed) and the CA dinucleotide (shaded) should be the bases removed from the right (U5) end prior to integration. A similar analysis should be possible for the left (U3) end of the molecule. However, the precise site of RNase H cleavage which defines the plus-strand primer is not known for all viruses. Where the site has been determined, it is always found within the PPT, two bases from the TG dinucleotide (Fig. 2, box) (20, 23, 25).

* Corresponding author.

FIG. 1. Outline of the reverse transcription and integration of the retroviral genome. The single-stranded RNA (ssRNA) genome (A) resembles a cellular mRNA in that it has a 3' poly(A) tail (A_n) and a 5' cap (Gppp). The genome contains short repeat sequences (R) and unique 3' and 5' sequences at each end (U3 and U5). The PPT and the tRNA PBS, where plus- and minus-strand DNA synthesis is initiated, are shown. Following reverse transcription, the dsDNA (B and D) is bounded by LTRs consisting of U3, R, and U5 sequences. In the nucleus the linear dsDNA molecule gives rise to two circular forms of dsDNA (C) containing one and two LTRs. IN cleaves a specific number of bases, usually two, from each 3' end of the linear dsDNA (E). The recessed 3' ends are covalently attached to the 5' ends of a double-stranded staggered cut in the host DNA. After DNA repair, the integrated provirus (F) is shorter than its double-stranded linear precursor and is flanked by a short direct repeat of host DNA. Additional details are given in the text (see also references 30 and 33).

Additional information that can be used to predict the ends of the linear DNA comes from analysis of the junction sequences in the two LTR circles found in the nucleus. Circles with two LTRs appear to be formed by blunt-end ligation of the linear dsDNA molecule (28, 30, 33). Whether

this ligation is catalyzed by cellular or viral enzymes is unclear. However, the bases that are lost from the ends of the linear viral DNA during integration are present at the circle junction in the circles with two LTRs.

We, and others, have reported previously on the se-

FIG. 2. Comparison of the viral genome/LTR boundaries and predicted circle junction sequences of several retroviruses. Rous sarcoma virus (Pr-C strain) (ALV), Moloney murine leukemia virus (MLV), and mouse mammary tumor virus (MMTV) sequences are from reference 33. HIV-1 sequence is from references 24, 27, and 32. HIV-2 sequence is from reference 13. The tRNA PBS and the PPT are shown in open boxes. The conserved CA/TG dinucleotides found at the boundaries of the integrated provirus are shown in shaded boxes. The site of RNase H cleavage of the PPT that produces the plus-strand primer is not known for HIV-2. The two most likely sites are designated by question marks. Our data (see Fig. 3) suggest that the cleavage site will be between the G and the A.

quences found at the circle junctions of human immunodeficiency virus type 1 (HIV-1) circles with two LTRs (15, 18, 22, 29, 34). These data suggest that integration is accompanied by the loss of two bases from each end, as for other retroviruses. This was contrary to what was predicted from cloned HIV-1 proviral sequences (Fig. 2) (24, 27, 32). Most retroviruses have two bases between the tRNA PBS (which ends GGT) and the CA dinucleotide found at the end of the integrated provirus. HIV-1 has only one base at this position. However, as we and others have shown, two bases from the U5 side are present in HIV-1 circle junctions (15, 18, 22, 29, 34), suggesting that the information contained in terminal A of the tRNA primer is found at the right end of the linear DNA molecule. We also presented evidence suggesting that the terminal ribo-A remains attached to the dsDNA molecule because the RNase H of HIV-1 reverse transcriptase does not remove the last A of the tRNA primer (34). These results led us to postulate that there was selective pressure that caused HIV-1, and perhaps all retroviruses, to maintain a substrate for integration from which two bases could be removed symmetrically from each end by IN.

However, the sequences of HIV-2 proviruses (13) suggest that retroviral INs may not always act symmetrically. As shown in Fig. 2, there are three bases between the end of the PBS and the conserved CA found at the right U5 boundary of the HIV-2 provirus. Even if HIV-2 RNase H removes the entire tRNA primer, three bases would be removed from the U5 end by IN prior to integration. There is no simple rule to predict how many bases will be lost from the U3 end. By analogy with other retroviruses, it could be predicted that two bases should be removed from the U3 side in preparation for integration (Fig. 2). Our results with HIV-1 suggest that HIV-1 IN acts symmetrically on both ends of the linear molecule. If this rule of symmetry is also true for HIV-2, three bases might be removed from the U3 side as well. The primer used for initiation of plus-strand synthesis in HIV-2 has not been precisely defined, and, depending on the precise site of RNase H cleavage, it is possible that either two or three bases (AC or GAC) would be removed from the U3 side of the linear molecule prior to integration. This would predict that five or six extra bases could be present at the circle junction of the circles containing two LTRs in HIV-2-infected cells.

HIV-2 (strain NIH-Z)-infected CEM cells harvested 24

and 48 h postinfection were kindly provided by David Waters. DNA was prepared by the method of Hirt (14), and the circle junction sequences were amplified by using primers in U3 and U5 (Fig. 3). Primer sequences containing *Eco*RI sites are as follows: U5, GCCGAATTCGGTCATTCGGTGTTCACCTG; U3, GCCGAATTCGTGTAGTTCTGCCAATCTGC. The polymerase chain reaction (PCR) was performed as follows. The final reaction mixture contained 50 mM Tris chloride (pH 8.3); 50 mM KCl; 7 mM MgCl₂; 170 μg of bovine serum albumin per ml; 1 mM each dATP, dCTP, dGTP, and dTTP; 1 μM each primers U3 and U5; 4% dimethyl sulfoxide; 1 μl of Hirt lysate; and 2 U of *Taq* polymerase (United States Biochemical Corp., Cleveland, Ohio) in a total volume of 25 μl. Thirty cycles were performed by using a Perkin-Elmer Cetus DNA thermal cycler with a step-cycle file programmed for 89°C for 40 s and 59°C for 80 s. The products of the amplification were fractionated on a 2% low-melting-temperature agarose gel and visualized by staining with ethidium bromide. The HIV-2-specific band (190 bp) was excised, extracted from the gel, and digested with *Eco*RI. The *Eco*RI fragment containing the circle junction was ligated into M13mp18 and used to transform competent *Escherichia coli* DH5α F' (Bethesda Research Laboratories). M13 clones with inserts were processed and sequenced by following the Sequenase DNA Sequencing Kit protocol (United States Biochemical Corp.).

We sequenced 30 circle junction clones; their sequences are summarized in Fig. 3B. The majority of the clones (24 of 30) contain five extra bases (GGTAC), three bases from U5 and two bases from U3. We also found several variant circle junction clones. One type of clone lacked the terminal T from the U5 side, and another lacked the terminal two bases (AC) from the U3 side. This type of aberrant circle junction has been described previously in reports on the sequence of HIV-1 circle junctions (18, 29). In two clones, seven nucleotides from U3 and 15 nucleotides from U5 were missing from the circle junction. We also found one clone in which the terminal base was lost from each end concomitant with the insertion of one extra base. It is not clear from the sequence exactly how this circle junction was formed. None of the clones we sequenced contained six extra bases at the circle junction, as would be expected if three bases were removed symmetrically from each end of the linear dsDNA prior to integration of HIV-2. The data also allow us to

FIG. 3. (A) Positions of primers (indicated by the arrows) in U3 and U5 used to amplify the circle junction sequences of HIV-2. (B) Sequences of HIV-2 circle junction clones. The conserved CA and TG dinucleotides found at the end of integrated HIV-2 proviruses are shown in shaded boxes. The sequence found most often is shown in the top line. The number in parentheses to the right of each clone is the number of clones seen with each sequence. Sequence identity is indicated by dashed lines. Point mutations in individual clones are indicated. (C) Sequence of a clone representing the right end of an autointegration event. The U5 boundary is indicated by an arrow.

predict the site of RNase H cleavage in the PPT of HIV-2. The cleavage must take place between the G and the A, two bases from the conserved TG found at the U3 boundary of the integrated HIV-2 provirus.

The 30 HIV-2 circle junction clones reported here were isolated following a single round of PCR amplification (30 cycles) with Hirt supernatant DNA as the substrate. We believe that most of the sequence diversity seen in these clones reflects differences in the population of circular HIV-2 DNAs used as a substrate for the amplification reaction. This diversity probably arose during reverse transcription of the retroviral DNA, although we cannot rule out the possibility that some of the sequence differences were introduced during PCR amplification. There are two clones that have an identical deletion in U3 and U5, and it is likely that these clones derive from the amplification of a single circle junction.

In addition to the circle junction clones, we report the sequences of clones derived from the amplification of one end of an autointegration event (Fig. 3C). When we originally attempted to clone the HIV-2 circle junction, we chose PCR primers that did not include *Eco*RI sites. Our attempts to clone the blunt-end fragments generated by PCR were unsuccessful (only one clone was isolated that contained HIV-2 circle junction sequences). However, when a sample of this material was used as a substrate for a second PCR with primers containing *Eco*RI sites, the amplified product was easily cloned into M13. Many of the clones, however, did not contain HIV-2 circle junctions but instead were found to represent one boundary from an autointegration event (Fig. 3C). These clones contain U5 sequences up to the CA residues found at the boundary of a normal integrated provirus. Examination of the adjacent sequences

shows that the site of integration is in the *env* region at base 7468. We were able to clone the PCR products from this integration event because a downstream sequence (bases 7557 to 7577) shares homology with the U3 primer chosen for PCR (an 18- of 21-base match). PCR with the U3 primer annealed in this position gives rise to a 210-bp fragment which was cloned along with the 190-bp circle junction-specific fragment.

These clones represent the right end of a provirus integrated into a target DNA, showing that integration in this particular strain is accompanied by a loss of three bases from U5. Ten clones were sequenced, all of which we assume were amplified from a single autointegrand. There are no sequence differences in these 10 clones in the region shown in Fig. 3C. However, two of the clones have independent point mutations outside this region. These clones were isolated following a total of 60 cycles of PCR, and it is likely that these mutations were introduced by the *Taq* polymerase during amplification. The relative lack of sequence diversity among these clones argues that the conditions we have used for PCR do not generate sufficient mutations to account for the diversity of circle junction sequences we have isolated.

We know of no other naturally occurring retrovirus in which the IN removes a different number of bases from the 3' and 5' ends prior to integration of the provirus. Goff and coworkers have constructed and analyzed a number of mutant Moloney murine leukemia viruses with changes at the right end of U5 (5, 6, 26). Some of these mutations result in U5 ends that are similar to the ends of HIV-1 and HIV-2 with respect to spacing and base content. Moloney murine leukemia virus mutants containing only one base between the PBS and the end of the provirus and mutants containing four bases at this position replicate reasonably well. It is still

unclear why HIV-1 uses a specific RNase H cleavage to generate a linear DNA substrate for integration with symmetric ends, whereas HIV-2 apparently generates an asymmetric substrate for integration. Although HIV-1 and HIV-2 are related viruses, the specificity of the RNase H and IN activities of these two viruses may differ. We are now attempting to use the specific substrates and enzymes in experiments to discover precisely what determines the specificity of these two reactions in HIV-1 and HIV-2.

We are grateful to David Waters and Larry Arthur for the gift of HIV-2-infected CEM cells, to Marilyn Powers for the oligonucleotides, and to Hilda Marusiodis for typing the manuscript.

Research was sponsored by the National Cancer Institute, under contract NO1-CO-74101 with ABL.

REFERENCES

- Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop. 1987. Correct integration of retroviral DNA *in vitro*. *Cell* **49**:347-356.
- 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.
- Bushman, F. D., and R. Craigie. 1990. Sequence requirements for integration of Moloney murine leukemia virus DNA *in vitro*. *J. Virol.* **64**:5645-5648.
- Bushman, F. D., T. Fujiwara, and R. Craigie. 1990. Retroviral DNA integration directed by HIV integration protein *in vitro*. *Science* **249**:1555-1558.
- Colicelli, J., and S. P. Goff. 1985. Mutants and pseudo revertants of Moloney murine leukemia virus with alterations at the integration site. *Cell* **42**:573-580.
- Colicelli, J., and S. P. Goff. 1988. Sequence and spacing requirements of a retrovirus integration site. *J. Mol. Biol.* **199**:47-59.
- Craigie, R., T. Fujiwara, and F. Bushman. 1990. The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro*. *Cell* **62**:829-837.
- Ellis, J., and A. Bernstein. 1989. Retrovirus vectors containing an internal attachment site: evidence that circles are not intermediates to murine retrovirus integration. *J. Virol.* **63**:2844-2846.
- Ellison, V., H. Abrams, T. Roe, J. Lifson, and P. Brown. 1990. Human immunodeficiency virus integration in a cell-free system. *J. Virol.* **64**:2711-2715.
- 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.
- Fujiwara, T., and K. Mizuuchi. 1988. Retroviral DNA integration: structure of an integration intermediate. *Cell* **54**:497-504.
- Goff, S. P. 1990. Integration of retroviral DNA into the genome of the infected cell. *Cancer Cells* **2**:172-178.
- Guyader, M., M. Emerman, P. Sonigo, F. Clavel, L. Montagnier, and M. Alizon. 1987. Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature (London)* **326**:662-669.
- Hirt, B. 1967. Selective extraction of polyoma DNA from infected mouse cell cultures. *J. Mol. Biol.* **26**:365-371.
- Hong, T., K. Drlica, A. Pinter, and E. Murphy. 1991. Circular DNA of human immunodeficiency virus: analysis of circle junction nucleotide sequences. *J. Virol.* **65**:551-555.
- Katz, R. A., G. Merkel, J. Kulkosky, J. Leis, and A. M. Skalka. 1990. The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. *Cell* **63**:87-95.
- 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.
- Kulkosky, J., R. A. Katz, and A. M. Skalka. 1990. Terminal nucleotides of the pre-integrative linear form of HIV-1 DNA deduced from the sequence of circular DNA junctions. *J. Acquired Immune Defic. Syndr.* **3**:852-858.
- Lobel, 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.
- Luo, G. X., L. Sharmeen, and J. Taylor. 1990. Specificities involved in the initiation of retroviral plus-strand DNA. *J. Virol.* **64**:592-597.
- 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.
- Pauza, C. D. 1990. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. *Virology* **179**:886-889.
- Pullen, K. A., and J. J. Champoux. 1990. Plus-strand origin for human immunodeficiency virus type 1: implications for integration. *J. Virol.* **64**:6274-6277.
- Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starich, 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**:277-294.
- Ratray, 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.
- 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.
- Sanchez-Pescador, R., M. D. Power, P. J. Barr, K. S. Steimer, M. M. Stempien, S. L. Brown-Shimer, W. W. Gee, A. Renard, A. Randolph, J. A. Levy, D. Dina, and P. A. Luciw. 1985. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science* **227**:484-492.
- Shank, P. R., and H. E. Varmus. 1978. Virus-specific DNA in the cytoplasm of avian sarcoma virus-infected cells is a precursor to covalently closed circular viral DNA in the nucleus. *J. Virol.* **25**:104-114.
- Smith, J. S., S. Kim, and M. J. Roth. 1990. Analysis of long terminal repeat circle junctions of human immunodeficiency virus type 1. *J. Virol.* **64**:6286-6290.
- 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.
- Vora, A. C., M. L. Fitzgerald, and D. P. Grandgenett. 1990. Removal of 3'-OH-terminal nucleotides from blunt-ended long terminal repeat termini by the avian retrovirus integration protein. *J. Virol.* **64**:5656-5659.
- Wain-Hobson, S., P. Sonigo, O. Danos, S. Cole, and M. Alizon. 1985. Nucleotide sequence of the AIDS virus, LAV. *Cell* **40**:9-17.
- Weiss, R., N. Teich, H. E. Varmus, and J. Coffin (ed.). 1985. *RNA tumor viruses*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- Whitcomb, J. M., R. Kumar, and S. H. Hughes. 1990. Sequence of the circle junction of human immunodeficiency virus type 1: implications for reverse transcription and integration. *J. Virol.* **64**:4903-4906.