Sequence Analysis of the Human DNA Flanking Sites of Human Immunodeficiency Virus Type 1 Integration

SCOTT W. STEVENS† AND JACK D. GRIFFITH*

Department of Microbiology and Immunology and Lineberger Comprehensive Cancer Center,
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7295

Received 26 January 1996/Accepted 5 June 1996

Human immunodeficiency virus type 1 (HIV-1) tagged with the *Escherichia coli* *supF* gene has been used to clone integrated HIV-1 proviruses. Sequence analysis of the 600 to 800 bp of human DNA adjacent to 29 clones revealed a propensity for HIV-1 to integrate near the *Alu* class of human repetitive elements.

In the life cycle of human immunodeficiency virus type 1 (HIV-1), integration of the viral genome into the host DNA is required for the production of progeny virus (4). Much is known about the integration reaction, and numerous studies have described the general nature of retroviral integration site selection (2, 8, 9, 11, 12, 14, 17, 19; for a review, see reference 15). However, over lengths of several hundred base pairs relatively few direct sequence data regarding the host sequences flanking integration sites are available. In a previous study of eight cloned HIV-1 proviruses containing significant amounts of human flanking DNA (16), seven of the eight proviruses were found to contain either a human long interspersed nuclear element 1 (L1Hs) or an *Alu* element within 200 bp of the provirus. This suggested that a similar analysis of a larger set would be worthwhile.

Because it was difficult to obtain a large library of HIV-1 proviral clones by conventional methods, a strategy pioneered with murine retroviruses (6, 7, 13), using a tagged retrovirus, was employed. Here, by inserting the bacterial *supF* gene into the 3' long terminal repeat (LTR) of HIV-1, it was possible to recover proviruses, using a lambda vector containing amber mutations and *Escherichia coli* cells which lack *supF* sequences. The *supF* gene was inserted into the *PmlI* site in the U3 region of the 3' LTR of the plasmid pCADA (Ron Swanson, University of North Carolina at Chapel Hill), which contains a derivative of the HIV-1 strain NL4-3 (Fig. 1A). The *supF* sequence is duplicated during reverse transcription and becomes part of both the 5' and the 3' LTR in the provirus (Fig. 1B). This facilitated the cloning of whole proviruses and their associated flanking sequences as well as partial-length proviruses containing only a 5' or 3' LTR together with flanking human sequences. The *supF* gene was inserted into the *nef* gene, which in pCADA contains a 60-bp deletion. Neither the 60-bp *nef* deletion nor the insertion of the *supF* gene was found to interfere with the replication or integration of the *supF*-containing virus (data not shown). Transcription from the 5' LTR of HIV^{supF} was not compromised, as the *PmlI* site is upstream of the TATA box and the major transcription factor binding sites (Fig. 1C). Reverse transcriptase assays from HeLa cells transfected (1) with the HIV^{supF} construct showed that both the parent vector and the *supF*-containing vectors produce reverse transcriptase at levels >1,000 fold above background levels (data not shown). Reverse transcriptase activity,

however, is not an indication of infectivity, integration, or replication. Therefore, the virus from the transfection supernatant was used to infect CEMx174 cells, a cell line highly permissive for infection with pCADA, the plasmid carrying the HIV-1 strain from which HIV^{supF} is derived. HIV^{supF} produced viable virus in CEMx174 cells, as shown by its ability to form syncytia, at levels identical to those of virus produced from the pCADA construct.

Genomic DNA from human cells infected with HIV^{supF} was partially digested with *Sau3AI*, and 15- to 24-kb DNA fragments were cloned into lambda Charon 40A (3) and packaged into phage. *E. coli* MC1061 cells (*sup*⁰) were infected with the resulting recombinant phage, and those containing HIV-1 proviral DNA formed plaques. Fifty-two candidate clones con-

FIG. 1. Construction of the HIV^{supF} clone. (A) Map of pCADA, which contains a full-length, transcription-competent molecular clone of HIV-1 derived from the HIV-1 NL4-3 strain. The *PmlI* site in the HIV-1 3' LTR was used for the insertion of the *supF* gene. (B) Representation of an integrated HIV^{supF} provirus. The *supF* gene, which is duplicated into the 5' LTR during reverse transcription, is denoted by a black rectangle. (C) Enlargement of the site of the *supF* gene insertion. The *PmlI* site is contained in the *nef* gene and is upstream of the TATA box and the SP-1 and NF- κ B binding sites.

* Corresponding author. Phone: (919) 966-2151. Fax: (919) 966-3015. Electronic mail address: jdg@med.unc.edu.

† Present address: Division of Biology 147-75, California Institute of Technology, Pasadena, CA 91125.

TABLE 1. HIV^{supF} 5' and 3' partial clones: presence of L1Hs or *Alu* elements in the human flanking DNA

Clone	Clone type	Total (bp) DNA sequenced	Repetitive element(s) encountered ^a	Identity length(s) (bp) ^b	% Identity to consensus ^{b,c}
HIVS1	5'	702	<i>Alu</i>	193 ^d	86
HIVS2	5'	694			
HIVS3	5'	691			
HIVS4	3'	639			
HIVS5	3'	601			
HIVS6	5'	684	<i>Alu</i>	123 ^d	75
HIVS7	5'	685	<i>Alu</i>	300 ^d	72
HIVS8	5'	693	<i>Alu</i>	292	82
HIVS9	3'	647	<i>Alu</i> (2)	110 ^d , 179 ^d	92, 78
HIVS10	5'	681	L1Hs	386 ^d	72
HIVS11	5'	703	<i>Alu</i> (2)	120 ^d , 230	72, 80
HIVS12	5'	680	L1Hs, <i>Alu</i>	255 ^d , 297	71, 79
HIVS13	3'	611	<i>Alu</i>	312	78
HIVS14	5'	685			
HIVS15	5'	659	<i>Alu</i>	270 ^d	81
HIVS16	5'	613			
HIVS17	5'	709			
HIVS18	5'	709			
HIVS19	5'	716			
HIVS20	3'	713	<i>Alu</i>	315	82
HIVS21	3'	729	<i>Alu</i> (2)	273 ^d , 130 ^d	78, 72
HIVS22	3'	687	<i>Alu</i>	300	80
HIVS23	3'	721	<i>Alu</i>	310	80
HIVS24	3'	699	<i>Alu</i>	261 ^d	78
HIVS25	5'	726			
HIVS26	3'	733	<i>Alu</i> (2)	255 ^d , 190 ^d	82, 71
HIVS27	5'	721			
HIVS28	5'	717	<i>Alu</i>	300	75
HIVS29	5'	719	L1Hs, <i>Alu</i>	250 ^d , 222	73, 71

^a Values in parentheses are numbers of elements encountered.^b When more than one value is listed, if an L1Hs and an *Alu* element were encountered the first value refers to L1Hs and the second value refers to *Alu*.^c Percent identity relative to GenBank L1Hs consensus sequence humtnl22 or *Alu* consensus humalup7.^d Sequence continues in clone beyond the segment sequenced here.

taining HIV-1 sequences as shown were identified by Southern hybridization. Further analysis indicated that 23 were full-length clones with various amounts of human DNA on both sides and 29 contained partial HIV-1 genomes because of cleavage at a *Sau3AI* site within the proviral DNA. Clones termed 5' clones contain 5'-flanking sequences and various amounts of HIV-1 provirus but do not contain 3' flanking sequences. Similarly, 3' clones were also present. Clones HIVS22 and HIVS29 (Table 1) could not be definitively assigned as 5' or 3' clones by Southern blotting. Sequence analysis, however, showed that HIVS29 was a 5' clone and that HIVS22 was a 3' clone (Table 2; Fig. 2 and 3).

Bacteriophage lambda DNA from 29 HIV^{supF} clones containing partial HIV-1 genomes (19 5' clones and 10 3' clones) was prepared, and the first 600 to 800 bp was sequenced by the UNC automated DNA sequencing facility. The sequence data were compared with the sequences in GenBank by using the BLAST and FASTA (10) modules of the Genetics Computer Group program using the default parameters. Clones which showed homology to either L1Hs or *Alu* sequences were then tested directly against a consensus sequence (humtnl22 for L1Hs elements or humalup7 for *Alu* elements) to normalize all of the sequence identities and element polarities. A homology window of greater than 70% is considered reasonable for comparisons of repetitive elements within a species (5) and was adopted for these analyses.

This analysis (Fig. 2 and 3; Table 1) showed that 3 of the 29 clones (10%) contained L1Hs sequences in their flanking DNA. In HIVS29, the provirus had integrated into an L1Hs element. The other L1Hs elements were 252 bp (HIVS10) and 427 bp (HIVS12) away from the beginning of the proviral sequences (Fig. 2). In clone HIVS2, an open reading frame which is 100% identical to an expressed sequence of over 148 bp in GenBank (H11552) was found. Although the gene for this expressed sequence is uncharacterized, the open reading frame corresponds to a possible exon, on the basis of the presence of conserved intron-exon junction sequences. Analysis of the clones (Fig. 2 and 3; Table 1) showed that *Alu* elements were present in 59% of the isolates, with several instances

TABLE 2. Sequence features of human DNA flanking sites of HIV-1 integration

Clone	Clone type	Reference	5-bp TSD ^a	Presence ^b of:		% A or T
				L1	<i>Alu</i>	
HIVS1	5'	This study	CAGTA	—	+	47
HIVS2	5'	This study	GTCTT	—	—	58
HIVS3	5'	This study	CCCTT	—	—	60
HIVS4	3'	This study	TTAGC	—	—	67
HIVS5	3'	This study	GTTTC	—	—	48
HIVS6	5'	This study	TGATT	—	+	67
HIVS7	5'	This study	GGAGG	—	+	50
HIVS8	5'	This study	GAAGC	—	+	58
HIVS9	3'	This study	ACTAA	—	+	52
HIVS10	5'	This study	CTCCT	+	—	62
HIVS11	5'	This study	ATCAC	—	+	53
HIVS12	5'	This study	CTTTT	+	+	57
HIVS13	3'	This study	AGCAG	—	+	52
HIVS14	5'	This study	CTTCC	—	—	61
HIVS15	5'	This study	GCCTC	—	+	52
HIVS16	5'	This study	GGGTG	—	—	36
HIVS17	5'	This study	GAGAT	—	—	57
HIVS18	5'	This study	GAGAC	—	—	52
HIVS19	5'	This study	CCAGC	—	—	47
HIVS20	3'	This study	TGGAA	—	+	58
HIVS21	3'	This study	TTCCC	—	+	50
HIVS22	3'	This study	CACCC	—	+	46
HIVS23	3'	This study	GGCAC	—	+	54
HIVS24	3'	This study	ATAAC	—	+	58
HIVS25	5'	This study	GTTAT	—	—	64
HIVS26	3'	This study	CTAGG	—	+	56
HIVS27	5'	This study	GCTGC	—	—	56
HIVS28	5'	This study	GTACA	—	+	50
HIVS29	5'	This study	AGCTC	+	+	58
YU-6	NA ^c	16	GAACC	+	—	NA
SF2	NA	16	GAAGG	+	—	NA
SF162	NA	16	AGTAT	—	+	NA
SF73	NA	16	TTCCT	+	—	NA
pSG31B	NA	16	AAAAT	+	—	NA
pSG95x	NA	16	GTATC	—	+	NA
pBENN5	NA	16	ACATA	—	+	NA
pPV60B	NA	16	GGATA	—	—	NA
168.7 1	NA	20	GCATG	NA	NA	NA
320.2A 1.1	NA	20	TTTAC	NA	NA	NA
320.2A 1.2	NA	20	GATCA	NA	NA	NA
320.2A 2.1	NA	20	ATAAG	NA	NA	NA
320.3 4	NA	20	AAGAG	NA	NA	NA
320.3 7	NA	20	GTGTT	NA	NA	NA
320.3 1	NA	20	GTATT	NA	NA	NA
320.3 6	NA	20	GTAAG	NA	NA	NA
HXB2	NA	20	GTAGT	NA	NA	NA

^a Sequence of 5-bp target site duplication (TSD).^b —, absent; +, present.^c NA, not applicable.

FIG. 2. Flanking sequences of HIV^{supF} clones containing the 5' LTR. The 5'-flanking DNA in the HIV^{supF} clones was analyzed for the presence of *Alu*, L1s, and open reading frame sequences by using the GenBank database. Transcriptional orientations of repetitive elements are indicated by arrows above the elements. The amounts of HIV-1 and human sequences are uncharacterized for these isolates and vary from clone to clone (see text).

of multiple repetitive elements: L1s/*Alu* chimeras (HIVS12 and HIVS29) and *Alu/Alu* chimeras (HIVS9, HIVS11, and HIVS21). Of the 21 *Alu* elements encountered, 10 were in the same transcriptional orientation as the provirus and 11 were in

the opposite orientation. Each of the three L1s elements was in the opposite transcriptional orientation as the provirus.

Target site duplication sequences have previously been analyzed for nine clones (20), from which a loose consensus of GTPuNN was derived. Table 2 contains the individual target site duplication sequences for the 29 HIV^{supF} clones, the clones analyzed in our previous studies, and the nine sequences from Vink et al. (20). There appears to be a preference for G residues at the first position (48%), T residues at the second position (44%), A residues at the third position (42%), A or T residues at the fourth position (65%), and T or C residues at the fifth position (65%). This would suggest a consensus sequence of GTA(A/T)(T/C), an elaboration of the previous consensus.

In this study a molecular clone of HIV-1 which contains the *E. coli supF* gene was constructed to selectively clone both full-length and partial HIV-1 proviruses and their human flanking sequences. In 59% of the partial clones whose sequences are described in this work, a human *Alu* repetitive element was found within 600 bp of the provirus. The significance of this value depends strongly on knowledge of the average spacing of *Alu* elements in the human genome. Many studies have suggested that *Alu* elements are clustered into "Alu islands" (17, 18). If so, and if HIV-1 preferentially integrates into *Alu* islands, this might suggest that these are regions of chromatin structure that are particularly favorable for both retroviral and retroposon integration. Further analysis of greater lengths of the flanking DNA in the clones described here and in additional examples should shed light on this possibility.

We are grateful to Brian Seed, Frederick Blattner, and Ron Swanson for their gifts of DNA and cells, Steve Pettit for his helpful conversations and help with the reverse transcriptase assays, and Yuh Hwa Wang, Sherry Stanfield-Oakley, and Terri Smith for their helpful conversations and technical assistance.

This work was funded by a grant from the NIH (GM 42342).

REFERENCES

1. Chen, C., and H. Okayama. 1987. High-efficiency transformation of mammalian cells by plasmid DNA. *Mol. Cell. Biol.* 7:2745-2752.

FIG. 3. Flanking sequences of HIV^{supF} clones containing the 3' LTR. The 3'-flanking DNA in the HIV^{supF} clones was analyzed for the presence of *Alu*, L1s, and open reading frame sequences by using the GenBank database. Transcriptional orientations of repetitive elements are indicated by arrows above the elements. The amounts of HIV-1 and human sequences are uncharacterized for these isolates and vary from clone to clone (see text).

2. **De Ambrosis, A., I. Casciano, F. Querzola, G. Vidali, and M. Romani.** 1992. Chromatin structure, DNA methylation, and gene expression at sites of viral integration in human fibroblasts—implications for chromosomal fragility. *Cancer Genet. Cytogenet.* **60**:1–7.
3. **Dunn, I. S., and F. R. Blattner.** 1987. Charons 36 to 40: multi enzyme, high capacity, recombination deficient replacement vectors with polylinkers and polystuffers. *Nucleic Acids Res.* **15**:2677–2698.
4. **Englund, G., T. S. Theodore, E. O. Freed, A. Engelman, and M. A. Martin.** 1995. Integration is required for productive infection of monocyte-derived macrophages by human immunodeficiency virus type 1. *J. Virol.* **69**:3216–3219.
5. **Hutchison, C. A., III, S. C. Hardies, D. D. Loeb, W. R. Shehee, and M. H. Edgell.** 1989. LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome, p. 593–617. *In* D. E. Berg and M. M. Howe (ed.), *Mobile DNA*. American Society for Microbiology, Washington, D.C.
6. **King, W., M. D. Patel, L. I. Lobel, S. P. Goff, and M. C. Nguyen-Huu.** 1985. Insertional mutagenesis of embryonal carcinoma cells by retroviruses. *Science* **228**:554–558.
7. **Lobel, L. I., M. Patel, W. King, M. C. Nguyen-Huu, and S. P. Goff.** 1985. Construction and recovery of viable retroviral genomes carrying a bacterial suppressor transfer RNA gene. *Science* **228**:329–332.
8. **Milot, E., A. Belmaaza, E. Rossart, and P. Chartrand.** 1994. Association of a host DNA structure with retroviral integration sites in chromosomal DNA. *Virology* **201**:408–412.
9. **Muller, H.-P., and H. E. Varmus.** 1994. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**:4704–4714.
10. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**:2444–2448.
11. **Pruss, D., F. Bushman, and A. Wolffe.** 1994. HIV integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci. USA* **91**:5913–5917.
12. **Pryciak, P. M., A. Sil, and H. E. Varmus.** 1992. Retroviral integration into minichromosomes *in vitro*. *EMBO J.* **11**:291–303.
13. **Reik, W., H. Weiher, and R. Jaenisch.** 1985. Replication-competent Moloney murine leukemia virus carrying a bacterial suppressor tRNA gene: selective cloning of proviral and flanking host sequences. *Proc. Natl. Acad. Sci. USA* **82**:1141–1145.
14. **Rohdewohld, H., H. Weiher, W. Reik, R. Jaenisch, and M. Breindl.** 1987. Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. *J. Virol.* **61**:336–343.
15. **Sandmeyer, S., L. Hansen, and D. Chalker.** 1990. Integration specificity of retrotransposons and retroviruses. *Annu. Rev. Genet.* **24**:491–518.
16. **Stevens, S. W., and J. D. Griffith.** 1994. Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1s repetitive elements. *Proc. Natl. Acad. Sci. USA* **91**:5557–5561.
17. **Stoppa-Lyonnet, D., P. E. Carter, T. Meo, and M. Tosi.** 1990. Clusters of intragenic alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. *Proc. Natl. Acad. Sci. USA* **87**:1551–1555.
18. **Timms, K. M., F. Lu, Y. Shen, C. A. Pierson, D. M. Muzny, Y. Gu, D. L. Nelson, and R. A. Gibbs.** 1995. 130 kb of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2-sulfate sulfatase locus. *Genome Res.* **5**:71–78.
19. **Vijaya, S., D. L. Steffen, and H. Robinson.** 1986. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J. Virol.* **60**:683–692.
20. **Vink, C., M. Groenink, Y. Elgersma, R. A. M. Fouchier, M. Tersmette, and R. H. A. Plasterk.** 1990. Analysis of the junctions between human immunodeficiency virus type 1 proviral DNA and human DNA. *J. Virol.* **64**:5626–5627.