Host sequences flanking the HIV provirus

Karen A.Vincent*, Deborah York-Higgins[1], Margarita Quiroga[1] and Patrick O.Brown
Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305 and
[1]Chiron Corporation, Emeryville, CA 94608, USA

Received June 26, 1990; Revised and Accepted September 12, 1990 EMBL accession no. X54195

ABSTRACT

A conserved property of retroviral proviruses is the presence of a direct repeat in the host DNA immediately flanking the viral sequence; each virus generates a repeat with a characteristic length. By sequencing the viral/host DNA junctions from five HIV-1 proviral clones, we have confirmed that integration of HIV results in the generation of a five basepair direct repeat. A target sequence in uninfected host DNA was analyzed to establish that the five basepair sequence flanking the provirus was present only once prior to integration. Of the five proviruses examined, two were found to have integrated in known repetitive sequence elements of the human genome; one in a Line-1 element and a second in satellite DNA.

INTRODUCTION

Integration of a DNA copy of the viral genome within a chromosome of the host appears to be an essential step in the life cycle of the human immunodeficiency virus (HIV; 1) as it is for other retroviruses. The integration event occurs early in the viral life cycle, following cellular entry and viral DNA synthesis. The reaction is thought to be mediated by viral factors contained within a nucleoprotein complex derived from the extracellular particle (2,3). This complex migrates to the nucleus where the integration event takes place. The provirus, which is thereafter a stable component of the host genome, is transcribed to produce both genomic RNA molecules and mRNA's encoding viral proteins.

Restriction enzyme and DNA sequencing analyses have defined several structural characteristics shared by all retroviral proviruses (see 4 for review). The provirus is bounded by long terminal repeats (LTR's) and is colinear with unintegrated viral DNA. The proviral/host DNA boundary always occurs at the ends of the LTR's. For most retroviruses, two basepairs present at each end of the unintegrated DNA are lost on integration. HIV is a possible exception to this rule; inspection of the viral sequence at the primer binding site suggests that a single basepair might be trimmed from the LTR termini (5). Many sites within the host genome can serve as targets for integration, although there is some evidence for preferred sites (6); integrated proviruses are often associated with transcribed regions (7) and DNase I-hypersensitive sites within chromatin (8,9). Finally, integration generates short tandem repeats of host DNA immediately flanking the provirus; the length of the duplicated sequence is characteristic of each virus.

The length of the flanking repeat and properties of the target site have important implications for the mechanism of integration. For HIV, the length of the repeat has not been established. Separate reports of single proviruses describe a repeat of five (10) or seven basepairs (11); in the latter case, a repeat of four basepairs is also consistent with the observed sequence.

By obtaining sequence data from four additional HIV-1 proviruses, we have been able to confirm that the length of the direct repeat is five basepairs. HXB-2, an HIV-1_{IIIB} proviral clone that had been characterized previously (11), was reanalyzed and found to conform to the five basepair rule. In addition, analysis of one target region established that the five basepair repeat is present only once prior to integration. Of the five proviruses studied, two were found to have integrated within repetitive sequence elements found in the human genome; the first (HIV-1SF2) within a Line-1 element (12) and the second (HIV-1_{SF170}) within satellite DNA (13).

METHODS

Clones used

The HIV-1_{SF2} isolate has been described previously (14). HIV-1_{SF33} (15), HIV-1_{SF162}, and $_{SF170}$ proviral clones were obtained in a similar manner; genomic DNA from infected Hut78 cells was partially digested with *EcoRI*, size-selected, and inserted into an EMBL-4 bacteriophage λ vector. For SF33 and SF162, subclones containing approximately half of each viral genome were generated by cleavage at an internal *EcoRI* site and ligation into a pUC19 vector (15). λHXB-2, a molecular clone of HIV-1_{IIIB} (16) was obtained through the NIAID AIDS Research and Reference Reagent Program.

DNA Sequencing

Plasmid DNA for sequencing was isolated by a modification of the Birnboim and Doly (17) procedure and further purified by passage over a mini anion-exchange column (Quiagen). The HIV-1_{SF170} and λHXB-2 DNA's were sequenced directly from the λ clones. For SF170, 14 kb and 6.4 kb *KpnI* fragments were isolated that contained the right and left viral/host junctions,

* To whom correspondence should be addressed

respectively. A 22 kb *XhoI* fragment contained the right junction of λHXB-2 while the left junction was found within a 6.5 kb *NarI* fragment. These were purified by electroelution from agarose.

Oligonucleotides used for sequencing were 17 mers complementary to the U3 (5′-AGGGAAGTAGCCTTGTG-3′) and U5 (5′-CTAGAGATCCCTCAGAC−3′) regions of the LTR. For HXB-2, an oligonucleotide complementary to host DNA flanking the right LTR (5′-GTCAAGGCCTCTCACT-C−3′) was also used.

Double-stranded DNA's (both supercoiled plasmid and linear λ fragments) were sequenced by the dideoxy method using a modified T7 polymerase (Sequenase version 2.0, U.S. Biochemicals). Sequencing reactions were carried out according to manufacturers specification except the labelling mix was diluted 1: 3 and the termination reactions were allowed to proceed for 30 min.

Searches of GenBank were performed using the GCG sequence analysis software package (18).

Polymerase Chain Reaction

The polymerase chain reaction (PCR; 19) was performed with two primers (5′-TCTAGATCTTTCCTGCTTTC-3′; 5′-GAA-TTCAGGATTAAGAAACT-3′) complementary to host sequences flanking the HIV-1_{SF2} integration site. These amplified a DNA fragment 242 bp long using as template genomic DNA isolated from Hut78 cells, an established human T-cell line (14). In addition to the template DNA, the reaction included 100 pmol of each primer in the presence of 50mM KCl, 10mM Tris-Cl (pH8.3), 1 mM $MgCl_2$, 0.01% gelatin, 200μM each dNTP and 2.5 U Taq polymerase (Perkin-Elmer Cetus) in a total volume of 100 μl. The denaturation, annealing, and polymerization steps were performed at 94°C (1 min), 50°C (2 min) and 72°C (3 min), respectively. The reactions were allowed to proceed through thirty cycles.

The product of the PCR reaction was gel-purified and kinased (20). The DNA was then ligated to pUC19 that had been cleaved with *HincII* and treated with calf intestinal phosphatase (Boehringer Mannheim). These clones were sequenced as described above.

RESULTS

The two existing reports in the literature (10,11) differ as to the length of the direct repeat in host DNA flanking the HIV provirus. Figure 1, part A lists the host/viral junction sequences of the four additional HIV-1 proviral clones analyzed here. 70−200 basepairs of flanking DNA was sequenced on each side of the proviruses. In each case, integration of HIV formed a direct repeat of five basepairs in the host DNA. Though the host sequence at the junction differs for each clone, the viral sequence is constant. Part B of Figure 1 shows the junction region of the HXB-2 clone and compares it to the previously-reported sequence (11). A single G residue, immediately 3′ to the right LTR, is missing from the published version. This result was confirmed by analyzing both strands of the DNA in this region. With the restitution of the omitted G, HXB-2, which was formerly reported to show a seven basepair duplication, now conforms to the five basepair rule established by the other proviruses. The host sequences show no obvious homology to each other or to the LTR termini.

Using the host flanking sequence determined in the experiments above, primers were designed to amplify the HIV-1_{SF2} target site by PCR. DNA from Hut78 cells was used as the template for the amplification reaction because the HIV-1 proviral clones were isolated originally from this cell line. Surprisingly, each of the eleven clones analyzed had a unique sequence. A search of the GenBank database revealed the cause of the unexpected sequence heterogeneity. The HIV-1_{SF2} provirus integration site is 94% identical to a Line-1 (L1) element sequence (12). The different products of the PCR reaction, therefore, result from amplification of the family of L1 elements that reside in the human genome. A consensus sequence of the PCR products is compared to the sequence of the HIV-1_{SF2} host/viral junction in Figure 2. Although it is impossible to make an unambiguous comparison of the target site prior to and following integration due to the observed heterogeneity, this experiment does establish that the five basepair sequence, GAAGG, is present only once in the target DNA prior to integration of the HIV provirus. Furthermore, no gross rearrangement of host DNA proximal to the integration site was observed.

A

```
SF2     GGAAATAACGAAATGAAGG[TG - - - CA]GAAGGCAGAAATAAAGATG
SF33    CATTCAGGGTATGAATATA[TG - - - CA]ATATATAAATATATCTTTG
SF162   TCAATCATTTTATCAGTAT[TG - - - CA]AGTATTCTCTCAGGCTCTC
SF170   AAGCTGTGGGCTTGGGATT[TG - - - CA]GGATTCAGTTCTTGAAACA
ref.10          TGTAGTGGG[TG - - - CA]GTGGGTGAT
```

B

```
ref.11          TAGTAGT[TG - - - CA]TAGTAGT
HXB-2   GTACTACAAACTTAGTAGT[TG - - - CA]GTAGTAGTTCATGTCATCT
```

Figure 1. A five basepair repeat in host DNA flanks the HIV provirus. Shown in part A are the host sequences immediately flanking each of the four HIV-1 proviruses examined here as well as a published sequence (ref. 10). The five basepair repeats are in bold type. Proviral sequences are represented by the bracketed region; the two terminal nucleotides of each LTR are shown. Part B is a comparison of the HXB-2 flanking region sequence with the published version (ref.11). The G nucleotide missing from the published sequence is underlined.

```
CON GAATTCAGGATTAAGAAACTCACTCAAAACCGCTCAACTACATGGAAACTGAACAACCTGC
FL                                    A   C A             G

CON TCCTGAATGACTACTGGGTACATAACGAAATGAAGGCAGAAATAAAGATGTTCTTTGAAA
FL                      A

CON CCAAT--------GAGAACAAAGACACAACATACCAGAATCTCTGGGACACATTTAAAGCA
FL       ATGGTCTTT            T

CON GTGTGTAGAGGGAAATTTATAGCACTAAATGCCCACAAGAGAAAGCAGGAAAGATCTAGA
FL       A                    G
```

Figure 2. The five basepair sequence is present only once in host DNA prior to integration. PCR was used to amplify a 242 basepair region of the human genome surrounding the HIV-1_{SF2} proviral integration target site. Shown in the figure is a comparison of the consensus sequence (CON) derived from the sequence of 11 different clones. The clones differ by an average of 5.2% from the consensus. Shown below is the flanking sequence derived from sequencing the host/viral junction region of the proviral clone (FL); only those positions which differ from the consensus target site sequence are shown. Sequences corresponding to the PCR primers are underlined. The five base pair sequence that is duplicated following integration appears in bold type.

To determine if any of the remaining HIV-1 proviruses had integrated into previously-characterized regions of the human genome, an additional search of GenBank was performed. A portion (111bp) of the HIV-1_{SF170} proviral flanking sequence shows 82% identity with a region of human satellite DNA (13).

DISCUSSION

The HIV-1 provirus, like all other retroviral proviruses (4), is flanked by a short direct repeat of host DNA. The length of this duplication has been reported to be either five (10) or seven (11) basepairs. We have determined the sequence of four additional HIV-1 proviral/host junctions and all were found to contain a direct repeat of five basepairs. Correction of an error in the Starcich *et al.* (11) sequence altered the length of the repeat from seven to five basepairs.

As has been documented in a similar analysis of an RSV integration site (21), the five basepair sequence is present only once in DNA isolated from uninfected cells. Generation of the repeat is therefore a consequence of the integration event. In a model for retroviral integration (2, 22 23), joining of the viral DNA ends to the host DNA is energetically coupled to staggered cleavage of DNA at the target site. During this reaction, the 3′ ends of the viral DNA are linked to the corresponding 5′ ends of the target DNA. Repair of the resulting heteroduplex generates the flanking repeat. The length of the duplication is specific for each retrovirus, suggesting the involvement of a virally-encoded factor in the target DNA cleavage reaction.

As with previous analyses of retroviral integration sites, (21, 24, 25, 26) no consistent features were shared by the host sequences flanking the HIV-1 provirus. This result indicates that, for HIV-1 as well, integration can occur at many sites in the genome. The analysis of so few proviral/host junctions, of course, does not rule out the existence of preferred sites, as have been described for integration of RSV (6). Furthermore, the target sequences show no homology to the viral DNA, implying that the integration reaction probably does not involve homologous recombination. The lack of specificity suggests that the choice of an integration site depends not on primary sequence, but on other characteristics of the target. This hypothesis is supported by the observation that integrated proviruses are commonly found associated with DNase I hypersensitive sites and/or transcribed regions of the chromosome (7, 8, 9).

The host sequences flanking the HIV-1_{SF2} provirus show extensive homology to the L1 family of repetitive sequence elements. The integration site is located within the second open reading frame of the L1 element. The amount of sequence difference observed between the integration site and the published L1 sequences is consistent with the average level of divergence observed among L1 elements in a single mammalian genome (7%; 27). Also consistent with this degree of variation, the eleven host target region clones differ by an average of 5.2% from the consensus. Though these sequences are transcriptionally active (12), integration of a provirus within one member of the L1 family may simply reflect the abundance of these elements in the human genome (approximately 10^4-10^5 per genome; 28).

In conclusion, we have confirmed that the length of the direct repeat in host DNA flanking the HIV provirus is five basepairs. A duplication of the same length is obtained *in vitro* when HIV DNA from an infected cell extract is reacted with an exogenous target (29). These results form a framework for the biochemical analysis of the integration reaction and may eventually lead to an understanding of how the HIV integration apparatus recognizes target DNA.

ACKNOWLEDGEMENTS

We thank Russell Higuchi for advice on the PCR and Vincent Young for help with the computer searches. This work was supported by grant AI27205 from the NIAID and by the Howard Hughes Medical Institute. P.B. is an assistant investigator of the Howard Hughes Medical Institute.

REFERENCES

1. Stevenson,M., Haggerty,S., Lamonica,C.A., Meier,C.M., Welch,S.-K. and Wasiak,A.J. (1990) J.Virol., 64, 2421–2425.
2. Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,J.M. (1987) Cell, 49, 347–356.
3. Bowerman,B., Brown,P.O., Bishop,J.M. and Varmus,H.E. (1989) Genes Dev., 3, 469–478.
4. Varmus,H.E. and Brown,P.O. (1989) In Howe,M. and Berg,D. (eds.), Mobile DNA. American Society for Microbiology Publications, Washington, D.C. pp.53–107.
5. Van Beveren,C., Coffin,J. and Hughes,S. (1985) In Weiss,R., Teich,N., Varmus,H. and Coffin,J. (eds.), RNA Tumor Viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., Vol.II, pp.359–1221.
6. Shih,C.-C., Stoye,J.P. and Coffin,J.M. (1988) Cell, 53, 531–537.
7. Scherdin,U., Rhodes,K. and Breindl,M. (1990) J.Virol., 64, 907–912.
8. Vijaya,S., Steffen,D.L. and Robinson,H.L. (1986) J.Virol., 60, 683–692.
9. Rohdewohld,H., Weiher,H., Reik,W., Jaenisch,R. and Breindl,M. (1987) J.Virol., 61, 336–343.
10. Muesing,M.A., Smith,D.H., Cabradilla,C.D., Benton,C.V., Lasky,L.A. and Capon,D.J. (1985) Nature, 313, 450–458.
11. Starcich,B., Ratner,L., Josephs,S.F., Okamoto,T., Gallo,R.C. and Wong-Staal,F. (1985) Science, 227, 538–540.
12. Skowronski,J., Fanning,T.G. and Singer,M.F. (1988) Mol.Cell.Biol., 8, 1385–1397.
13. Frommer,M., Prosser,J. and Vincent,P.C. (1984) Nucleic Acids Res., 12, 2887–2900.
14. Luciw,P.A., Potter,S.J., Steimer,K., Dina,D. and Levy,J.A. (1984) Nature, 312, 760–763.
15. York-Higgins,D., Cheng-Mayer,C., Bauer,D., Levy,J.A. and Dina,D. J.Virol., in press.
16. Shaw,G.M., Hahn,B.H., Arya,S.K., Groopman,J.E., Gallo,R.C. and Wong-Staal,F. (1984) Science, 226, 1165–1171.
17. Birnboim,H.C. and Doly,J. (1979) Nucleic Acids Res., 7, 1513–1523.
18. Devereux,J., Haeberli,P. and Smithies,O. (1984) Nucleic Acids Res., 12, 387–395.
19. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Science, 239, 487–491.
20. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor University Press, Cold Spring Harbor.
21. Hughes,S.H., Mutschler,A., Bishop,J.M. and Varmus,H.E. (1981) Proc.Natl.Acad. Sci., 78, 4299–4303.
22. Fujiwara,T. and Mizuuchi,K. (1988) Cell, 54,497–504.
23. Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,J.M. (1989) Proc.Natl. Acad.Sci., 86, 2525–2529.
24. Dhar,R., McClements,W.L., Enquist,L.W. and VandeWoude,G.F. (1980) Proc.Natl.Acad.Sci., 77, 3937–3941.
25. Hughes,S.H., Shank,P.R., Spector,D.H., Kung,H.-J., Bishop,J.M., Varmus,H.E., Vogt,P.K. and Breitman,M.L. (1978) Cell, 15, 1397–1410.
26. Shimotohno,K and Temin,H.M. (1980) Proc.Natl.Acad.Sci., 77, 7357–7361.
27. Soares,M.B., Schon,E. and Efstradiatis,A. (1985) J.Mol.Evol., 22, 117–133.
28. Singer,M.F. and Skowronski,J. (1985) Trends Biochem.Sci., 10, 119–122.
29. Ellison,V., Abrams,H., Roe,T., Lifson,J. and Brown,P. (1990) J.Virol., 64, 2711–2715.