JOURNAL OF VIROLOGY, May 1998, p. 4005–4014 Vol. 72, No. 5
0022-538X/98/$04.00+0
Copyright © 1998, American Society for Microbiology

Chromosome Structure and Human Immunodeficiency Virus Type 1 cDNA Integration: Centromeric Alphoid Repeats Are a Disfavored Target

SANDRINE CARTEAU, CHRISTOPHER HOFFMANN, AND FREDERIC BUSHMAN*

Infectious Disease Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037

Received 22 August 1997/Accepted 19 January 1998

Integration of retroviral cDNA into host chromosomal DNA is an essential and distinctive step in viral replication. Despite considerable study, the host determinants of sites for integration have not been fully clarified. To investigate integration site selection in vivo, we used two approaches. (i) We have analyzed the host sequences flanking 61 human immunodeficiency virus type 1 (HIV-1) integration sites made by experimental infection and compared them to a library of 104 control sequences. (ii) We have also analyzed HIV-1 integration frequencies near several human repeated-sequence DNA families, using a repeat-specific PCR-based assay. At odds with previous reports from smaller-scale studies, we found no strong biases either for or against integration near repetitive sequences such as *Alu* or LINE-1 elements. We also did not find a clear bias for integration in transcription units as proposed previously, although transcription units were found somewhat more frequently near integration sites than near controls. However, we did find that centromeric alphoid repeats were selectively absent at integration sites. The repeat-specific PCR-based assay also indicated that alphoid repeats were disfavored for integration in vivo but not as naked DNA in vitro. Evidently the distinctive DNA organization at centromeres disfavors cDNA integration. We also found a weak consensus sequence for host DNA at integration sites, and assays of integration in vitro indicated that this sequence is favored as naked DNA, revealing in addition an influence of target primary sequence.

To replicate, a retrovirus must integrate a cDNA copy of its RNA genome into a chromosome of the host. The host integration acceptor sites are not expected to be present as naked DNA but rather associated with histones and other DNA-binding proteins in chromatin. DNA packaging in vivo is expected to influence integration site selection, and the choice of integration site may have profound effects on both the virus and the host (13, 57). The determinants of integration efficiency in vivo remain incompletely defined, despite their importance.

Previous surveys of in vivo integration sites have led to several proposals for factors influencing site selection. Studies of Moloney murine leukemia virus have supported a model in which open chromatin regions at transcription units were favored, since associated features such as DNase I-hypersensitive sites (45, 58) or CpG islands (47) were apparently enriched near integration sites. Another study proposed that unusual host DNA structures were common near integration sites (34). A recent study of avian leukosis virus integration frequencies at several chromosomal sites failed to show any major differences among the regions studied (62), contrary to an earlier report (50). For human immunodeficiency virus type 1 (HIV-1), it has been proposed that integration may be favored near repetitive elements (including LINE-1 elements [54] or *Alu* islands [55]) or topoisomerase cleavage sites (24).

Assays of integration in vitro have revealed several effects of proteins bound to target DNA. Simple DNA-binding proteins can block access of integration complexes to target DNA, creating regions refractory for integration (3, 9, 44). In contrast, wrapping DNA on nucleosomes can create hot spots for integration at sites of probable DNA distortion (40–42, 44). Distortion of DNA in several other protein-DNA complexes can also favor integration (3, 35), consistent with the possibility that DNA distortion is involved in the integrase mechanism (11, 48).

Here we present two experiments designed to address some of the questions surrounding integration site selection in vivo. We have (i) sequenced 61 integration junctions made after experimental infection of cultured human T cells and compared them with 104 control DNA fragments from uninfected human cells and (ii) used a region-specific PCR assay to assess the frequency of integration near several repeated-sequence families. In addition, we have identified a weakly conserved sequence at in vivo integration sites and determined that it is favored for integration when tested in vitro.

MATERIALS AND METHODS

DNA manipulation. Plasmids containing synthetic integration target sites were prepared by annealing pairs of oligonucleotides (CH10-1–CH10-2, CH11-1–CH11-2, and CH13-1–CH13-2) (Table 1) and ligating them with pUC19 DNA that had been cleaved with *Eco*RI and *Hin*dIII. The standard cloning methods used were as described previously (46). Integration target DNAs were prepared by cleaving the plasmids mentioned above with *Pvu*II, which releases the oligonucleotide insert together with flanking plasmid DNA.

The oligonucleotides used in this study are shown in Table 1.

Construction of DNA libraries. To generate a large pool of independent integration events, SupT1 cells (2×10^7 cells) were infected with the HXB2 or R9 (56) (referred to as R8 in reference 22) HIV-1 strain. Viral stocks were assayed by measuring the concentration of p24, and the infectivity was scored by the MAGI assay (28). Cells were infected at a multiplicity of 1 to 10 and harvested 12 to 14 h later. The cellular genomic DNA was depleted of low-molecular-weight DNA prior to cloning as described previously (39).

For construction of library 1 (Fig. 1, method 1), DNA from infected cells was cleaved with *Hin*dIII and circularized by ligation (31). Sixty-six nanograms of DNA was used as the template for PCR. HUA and HUB, divergently oriented primers complementary to the HIV long terminal repeats (LTRs), were used for the first amplification. Amplification was carried out for 35 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 3 min. The products were purified by using the Qiaquick PCR purification kit (Qiagen, Santa Clarita, Calif.). One microliter

* Corresponding author. Mailing address: Infectious Disease Laboratory, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd., La Jolla, CA 92037. Phone: (619) 453-4100, ext. 1630. Fax: (619) 554-0341. E-mail: rick_bushman@qm.salk.edu.

TABLE 1. Oligonucleotides used in this study

Oligonucleotide	Sequence	Comments
HUA	5′-CTTTTTGCCTGTACTGGGTCTC-3′	HIV U3 primer for inverse PCR
HUB	5′-GATCAAGGATATCTTGTCTTCGT-3′	HIV U3 primer for inverse PCR
IP3	5′-TCTTGTCTTCGTTGGGAGTGA	HIV U3 primer for inverse PCR
det3b	5′-GAACCCACTGCTTAAGCCTC-3′	HIV U3 primer for inverse PCR
det3a	5′-CTTCGTTGGGAGTGAATTAG-3′	Primer for detection of circle junctions
sc8	5′-CTTCAAGTAGTGTGTGCCCG-3′	Primer for detection of circle junctions
sc10	5′-GGGTTTTCCAGTCACACCTCAGG-3′	Primer for detection of the HIV internal fragment
TA6	5′-CATCAAGCTTGGTACCGAGC-3′	Primer for sequencing from pTA vector
TA7	5′-TAATACGACTCACTATAGGG-3′	Primer for sequencing from pTA vector
SC24	5′-TGGCGCAATCTCGGCTCAC-3′	Primer for amplifying *Alu*1 sequences
CH12	5′-CTCCGCTTCCCGGGTTC-3′	Primer for amplifying *Alu*1 sequences
CH5	5′-CTTCCAGTTTTTGCCCATTCAGT-3′	Primer for amplifying LINE-1 sequences
CH6	5′-AGTATGATATTGGCTGTGGGTTTGTC-3′	Primer for amplifying LINE-1 sequences
SC21	5′-GCAAGGGGATATGTGGACC-3′	Primer for amplifying alphoid repeats
SC23	5′-ACCACCGTAGGCCTGAAAGCAGTC-3′	Primer for amplifying alphoid repeats
CH15	5′-CCTGAGGCCTCCCTCAGCCAT-3′	Primer for amplifying THE 1 repeats
CH16	5′-GCCATGATTGTAAGTTTCCTGAGG-3′	Primer for amplifying THE 1 repeats
NEB-40	5′-GTTTTCCCAGTCACGAC-3′	Primer for amplifying integration products in pUC19
FB652	5′-TGTGGAAAATCTCTAGCA-3′	Primer for amplifying HIV U5 sequences
CH 11	5′-CTCCGCTTCCCGGGTTC-3′	Primer for amplifying integration products in pUC19
FB66	5′-GCCTAGATCCGTGTGGAAAATC-3′	Primer for amplifying products made with purified integrase
FB64	5′-ACTGCTAGAGATTTTCCACACGGATCCTAGGC-3′	Substrate for purified integrase (annealed to FB65-2)
FB65-2	5′-GCCTAGGATCCGTGTGGAAAATCTCTCTCTAGCA-3′	Substrate for purified integrase (annealed to FB64)
AP1	5′-CCATCCTAATACGACTCACTATAGGGC-3′	Adaptor primer 1
AP2	5′-ACTCACTATAGGCTCGAGCGGC-3′	Adaptor primer 2
ADAPT1	5′-CTAATACGACTCACTATAGGGCTCGAGCGGCCGCCCGGGCAGGT-3′	Vectorette adaptor primer (top strand)
ADAPT2	5′-ACCTGCCC-NH2-3′	Vectorette adaptor primer (bottom strand)
CH10-1	5′-AATTCTTCTCGAGTAGGTTACCTATGATCAA-3′	Insert for pCH10 (top strand)
CH10-2	5′-AGCTTTGATCATAGGTAACCTACTCGAGAAG-3′	Insert for pCH10 (bottom strand)
CH11-1	5′-AATTCTTCTCGAGTAGTTTAACTATGATCAA-3′	Insert for pCH11 (top strand)
CH11-2	5′-AGCTTTGATCATAGTTAAACTACTCGAGAAG-3′	Insert for pCH11 (bottom strand)
CH13-1	5′-AATTCGTGTTAACTCGGTGACCGAAGGCCTA-3′	Insert for pCH12 (top strand)
CH13-2	5′-AGCTTAGGCCTTCGGTCACCGAGTTAACACG-3′	Insert for pCH12 (bottom strand)

from the 50-μl column eluate was used as the template for the second-round PCR (20 cycles; program as described above) with nested primers det3b and IP3.

For construction of library 2 (Fig. 1, method 2) DNA fragments sheared by sonication (average length, about 1.5 kb) were made blunt-ended by treatment with *Bal* 31 followed by T4 DNA polymerase and deoxynucleoside triphosphates. Ligation of adapters, amplification, and cloning were carried out as described previously (51), except that primers HUB and IP3 were used as viral end primers for the first and second amplifications, respectively. PCR products were cloned by using the pCR II TA cloning vector from Invitrogen (San Diego, Calif.).

The products of PCRs contained two contaminants in addition to the desired integration junctions, one derived from a circular form of the viral DNA (2-LTR circle) and the second from the 3′ internal part of the viral DNA (for a discussion, see reference 31). Colonies containing host-virus junctions were distinguished from colonies containing contaminating sequences by PCR. Bacterial colonies containing plasmids were resuspended in PCR buffer and amplified with *Taq* polymerase for 20 cycles of 1 min at 94°C, 30 s at 60°C, and 1 min at 72°C. The circle junctions were detected using primers det3a and sc8. The internal fragment was detected using primers sc10 and IP3. The inserts were sequenced by using primers TA6 and TA7, which are complementary to the vector (pCR II; Invitrogen). Sequences of integration junctions and controls were determined by the dideoxy sequencing method.

Each sequence was determined at least twice. For each integration site clone, the sequence of 34 bases of viral DNA at the LTR tip was determined, in addition to the flanking host DNA. For most integration site clones (59 of 61), all of the cloned human DNA adjacent to the proviral DNA was sequenced.

A control experiment was carried out to exclude a possible artifact. Since DNA samples were treated with DNA ligase, free HIV genomes might have become joined to host DNA fragments by DNA ligase instead of integration. This is unlikely in the case of library 1, however, since the blunt-ended or 3′ cleaved forms of the HIV cDNA would not be expected to become ligated to the protruding 5′ ends generated by cleavage with *Hin*dIII. However, to document this expectation, a control experiment was performed in which purified unintegrated HIV cDNA was incubated in the presence of DNA ligase with *Hin*dIII-cleaved sequences and possible ligation was assayed by PCR across the ligation junction (one primer complementary to the HIV DNA and the other complementary to the *Hin*dIII-cleaved test DNA). No ligation was detected (data not shown). In the case of library 2, hypothetical ligation of unintegrated HIV cDNA should have yielded predominantly the vectorette linker joined directly to HIV cDNA, since DNA ends from the linkers were present in vast excess over ends from viral or human DNA. However, no such forms were detected (data not shown). Internal evidence also argues against this class of artifacts. For example, the 5-bp consensus host sequence flanking integration sites identified here closely resembles that found in a previous study employing conventional cloning and sequencing (55), an observation that helps validate each study.

DNA sequence analysis. Sequences were analyzed by comparison to the nonredundant human sequence (nr) database, the human cDNA (dbEST) database, and the MONTH (November 1997) database by using BLASTN with Search Launcher and Repeat Masker. Default parameters were used. For comparisons between integration sites and control libraries, only a subset of the available sequence was considered (see Table 2), with either an average length of 144 bp or a length of exactly 50 bp (see Table 3). A total of 8,809 bp of human DNA flanking 61 integration sites was sequenced and analyzed for the integration site libraries (see Tables 2 and 3). The lengths of flanking human DNA sequences analyzed ranged from 37 to 430 bp. For the control human DNA fragments, a total of 14,989 bp in a total of 104 DNA clones were sequenced. Lengths of sequences analyzed ranged from 51 to 264 bp. Links to integration site and control sequences can be found at http://www.salk.edu/faculty/bushman.html.

Similarities to repeated sequences were ranked in accordance with the Smith-Waterman parameter (SW) generated by Repeat Masker (see A. F. A. Smit and P. Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) or by the probability of matching by chance generated by BLASTN (1) (*P* value) (see http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-blast?Jform=0). Minimum similarities for each sequence class considered to be significant matches are as follows: cDNA, $P = 4.6 \times 10^{-6}$; LINE 1, SW = 217; *Alu* repeat, SW = 195; alphoid repeat, SW = 218; other repeats, SW = 190. Most regions of sequence similarity extended over at least 50 bp, although in the case of the lowest scoring cDNA, a 31-bp perfect match was judged to be significant.

Integration in vitro. Preintegration complexes (PICs) were extracted from a 6-h coculture of SupT1 cells grown in RPMI 1640 medium containing 10% fetal calf serum and chronically infected MoltIIIB cells stimulated with phorbol 12-myristate 13-acetate as previously described by Farnet and Haseltine (19). In vitro integration was achieved by incubating 400 μl of PIC extract with 1.2 μg of DNA from uninfected SupT1 cells for 45 min. The integration product was

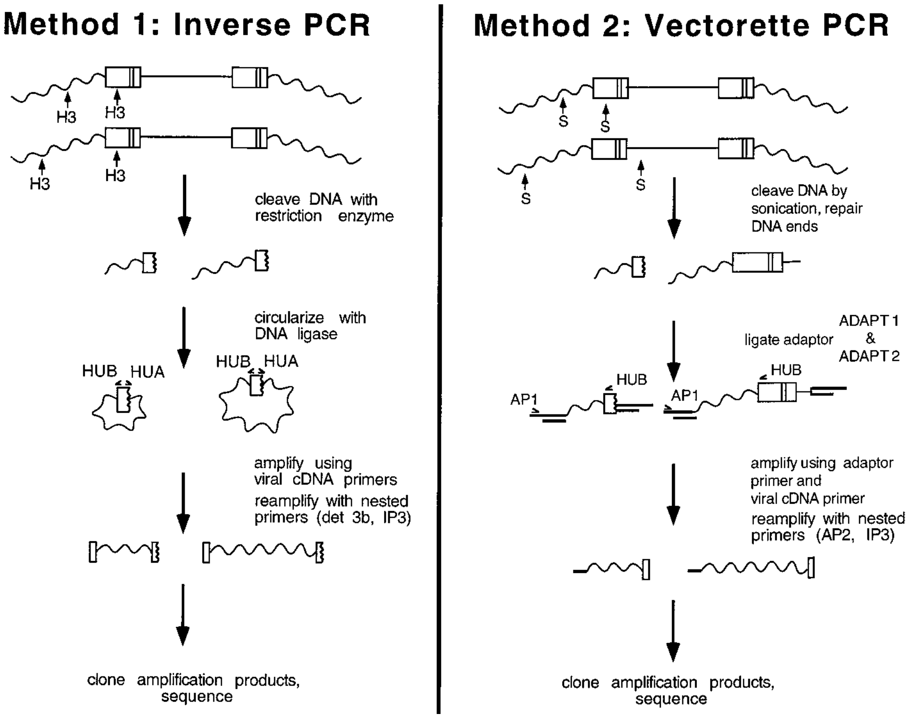

Method 1: Inverse PCR
H3
H3
H3
H3
cleave DNA with restriction enzyme
circularize with DNA ligase
HUB HUA
HUB HUA
amplify using viral cDNA primers
reamplify with nested primers (det 3b, IP3)
clone amplification products, sequence
Method 2: Vectorette PCR
S
S
S
S
cleave DNA by sonication, repair DNA ends
ADAPT1
ligate adaptor &
ADAPT2
AP1
HUB
AP1
HUB
amplify using adaptor primer and viral cDNA primer
reamplify with nested primers (AP2, IP3)
clone amplification products, sequence

FIG. 1. Cloning strategies for constructing integration site libraries. See the text for details and Table 1 for the sequences of oligonucleotides used.

recovered by incubating it with proteinase K in 0.5% sodium dodecyl sulfate followed by extraction with phenol-chloroform. The same procedure was followed for the inactive PICs after first incubating the concentrated PICs in 15 mM EDTA for 5 min prior to adding target DNA. Integration assays with recombinant HIV-1 integrase were carried out essentially as described previously (4, 10).

Region-specific analysis of integration acceptor sites. Integration junctions were amplified essentially as described previously (9, 30, 44). Cellular DNA templates were prepared from infected and uninfected samples as described above. Integration products were visualized by nested PCR. Products were first amplified with viral primer HUB and a repeat primer. Products were then reamplified with the viral primer IP3 which had been end labeled by treatment with [γ-^{32}P]ATP and kinase and a nested repeat primer. The primers for repeated sequences were designed by aligning multiple repeat copies and identifying conserved regions. Primers for amplifying repeated sequences were as follows (see Table 1 for sequences; in each case, the second primer is the nested second primer). *Alu*1, SC24 and CH12 (27); LINE-1, CH5 and CH6 (64); alphoid repeat, SC21 and SC23 (61); and THE 1, CH15 and CH16 (52). The amounts of integration products generated in vivo and in vitro that were used as templates for PCR were adjusted to provide equal numbers of proviruses in each case. The first round of PCR was carried out for 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min. For the second round of PCR, 2 μl from the initial PCR was added to a 25-μl reaction mixture and the mixture was amplified for 20 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 30 s. TaqStart antibody (Clontech, Palo Alto, Calif.) was used in both amplifications (hot-start PCR) in accordance with the manufacturer's recommendations.

Assays of integration into cloned target DNAs were carried out as described previously (for PICs [4, 8] and for purified integrase [3, 33]). PICs were concentrated and partially purified by pelleting through 20% sucrose as described before (4). Integration targets were (i) a purified *Pvu*II fragment containing the sequence of interest (PICs) or (ii) uncleaved plasmid DNA (purified integrase). Similar results were also obtained with PICs when uncleaved plasmid DNAs were used as the target. Primers for amplifying integration products were as follows: PIC reactions, top strand, NEB-40 and FB 652 (4); PIC reactions, bottom strand, CH 11 and FB 652; purified integrase reactions, top strand, FB 66 (4) and NEB-40; purified integrase reactions, bottom strand, FB 66 and CH 11.

RESULTS

Construction of integration site libraries. DNA for library construction was obtained from a human T-cell line (SupT1) acutely infected with cell-free stocks of HIV-1. Cellular DNA was harvested 12 to 14 h after initiation of infection, allowing initial integration to be studied separately from selection during subsequent growth of cells.

Libraries were constructed by two different methods in an effort to control for possible biases introduced in the DNA cloning steps (Fig. 1). For library 1, genomic DNA from infected cells was digested with *Hin*dIII, which cleaved the population of proviruses near the viral DNA ends and at numerous positions in flanking host DNA. *Hin*dIII-cleaved DNA was then circularized by treatment with DNA ligase, and virus-host DNA junctions were amplified with divergent primers complementary to viral end sequences (inverse PCR) (31, 49). For library 2, DNA fragments were made blunt ended by treatment with *Bal* 31 nuclease and T4 DNA polymerase and ligated to short linkers. DNA fragments were amplified with primers complementary to the linker and the HIV cDNA end (vectorette PCR) (51). PCR fragments were then cloned and sequenced. Sixty-one integration sites were analyzed by this means.

To aid in interpretation of the data, control libraries were constructed from uninfected SupT1 cell DNA by methods parallel to those used for cloning integration sites. SupT1 DNA fragments were generated by cleavage with *Hin*dIII (control library 1) or sonication and end repair (control library 2), cloned into plasmid vectors, and sequenced. One hundred four control clones from uninfected human DNA were characterized by this means.

Analysis of integration site libraries. Analysis of the sequencing data presented several challenges. Our raw sequence data contained different numbers of base pairs determined for each DNA clone analyzed. To compare the integration site and control data sets in a meaningful fashion, it was necessary to compare matching numbers of base pairs in each DNA clone and then compare the frequencies of appearance of different types of sequences in each data set. The average length of host DNA flanking integration sites was 144 bp, so sequences in the control library, which were slightly longer, were each truncated to yield test sequences with an average length of 144 bp (further parameters describing the data sets are presented in Materials and Methods).

Some copies of the human repeated DNA sequences are quite divergent from the family consensus sequence, presenting a challenge for identification. Repeated sequences were identified here by a two-step process. The program Repeat Masker, which compares unknown sequences to a set of consensus sequences derived from human repeat sequences (52), was used first. In a second step, all sequences were compared to the nr, dbEST, and MONTH (November 1997) databases by using BLASTN with default settings. In some cases, highly repeated sequences missed by Repeat Masker were identified by BLASTN and further analysis allowed them to be grouped into known sequence classes. The minimum degrees of similarity scored as matches are given in Materials and Methods.

Analysis of cDNA matches presented another challenge. New sequences are being added to the dbEST database at a high rate, and even during the course of this work many anonymous sequences were found in later searches to match new cDNAs. The data presented here represent the number of matches to cDNAs as of November 1997, but new additions to the database will likely increase the number of matches in the future. For cDNAs, there was a natural partitioning of sequences into plausible and unlikely matches, since integration into a transcribed region should yield a near-perfect match over a discrete region.

Integration sites sequenced and the matches to known sequences are summarized in Table 2 and 3. Sequences were classified as transcription units, *Alu* elements, LINE elements, alphoid repeats, other repeats, or anonymous. Transcription units were identified in database searches either as cDNAs or as sequences within the transcribed regions of known genes. *Alu* elements and LINE elements are the familiar interspersed nuclear repeats characteristic of human DNA. Alphoid repeats comprise the alpha satellite DNA, tandem arrays of 171-bp repeats associated with centromeric heterochromatin (38, 61). The "other repeat" class included several types, namely, SINE elements apart from *Alu* elements, low-complexity repeats, and retrovirus-related sequences such as THE 1 elements (36) and MLT1 sequences (14, 52) (for a recent summary of nomenclature, see reference 52). Anonymous sequences were defined as sequences contained in none of the classes.

For the control libraries, *Alu* sequences were identified in 10% of clones. Previous studies suggest that *Alu* elements comprise 8 to 15% of the human genome (53). LINE-1 elements comprised 13% of the control sequences; 5 to 18% was expected (16, 25, 53). Information available on transcription units, alphoid repeats, and the other repeats was insufficient to allow their abundance to be predicted with confidence. Analysis of the %GC of DNA in control library clones and in human DNA flanking integration sites revealed no obvious differences from that of bulk human DNA (data not shown). Thus, in those cases that could be checked, sequences in our control libraries had compositions close to those expected for randomly selected human genomic DNA fragments.

Comparison of the integration site and control libraries revealed that centromeric alphoid repeats were absent among integration sites but that six alphoid repeats were present in the control libraries (Tables 2 and 3). Alphoid repeats were also absent among previously characterized HIV-1 integration sites (37, 59).

Other types of sequences were differentially distributed between integration site sequences and control sequences, although none showed the all-or-nothing partitioning characteristic of alphoid repeats. Transcription units were more abundant in the integration sites (18%) than in controls (8%). The other repeats were also differentially distributed (7%) in integration sites versus 23% in controls), although in this case many different sequence types contributed to the totals. *Alu* elements and LINE elements were not obviously differentially distributed.

As a test of the robustness of our conclusions, integration site sequences were reanalyzed after truncation so that only 50 bp of host DNA remained at the junction between viral and host sequences for all clones. The control data was similarly truncated to 50 bp in each sequence, arbitrarily starting from one junction with the DNA vector used for cloning. Sequence similarities were identified in the 50-bp data set by using the criteria described above (Table 3). Fewer matches were detected, as expected, since the sequences were shorter. However, in this case also, alphoid repeats were detected in the control library and not the integration site library.

A weak consensus sequence at integration sites. Figure 2 presents an analysis of the 5 bp of host DNA at the junction between virus and host sequences expected to be duplicated upon integration. A weak consensus sequence can be derived from this data [5′ GT(A/T)AC 3′]. Only one end was sequenced for each integrant, so the duplicated nature of this sequence is inferred. The consensus sequence is rotationally symmetric, as expected, since each end of the HIV cDNA is joined to the 5′ end of each strand of this sequence (Fig. 2). A closely related sequence was derived from a previous study of HIV integration sites by Stevens and Griffith [5′ GTA(A/T)(T/C) 3′] (55). In this study, DNA from HIV-infected cells was cloned in lambda vectors, followed by isolation of provirus-containing clones by hybridization and sequencing of 29 proviral integration sites. The observation that our methods and that of Stevens and Griffith yielded similar integration site consensus sequences strongly validates each study.

Region-specific assays of integration target sites. Several features of the sequencing data complicated interpretation. (i) The number of matching sequences detected was determined in part by the choice of parameters in the similarity search. (ii) In some clones the integration junctions were within the identified cDNA or repeated sequence, while in others the junctions were near but not within the identified sequence. In Tables 2 and 3, these were considered together. (iii) Although this study of HIV-1 integration site sequences is the largest yet reported, the differences between integration sites and controls were generally not clearly significant, as evaluated by the chi-square or Fisher's exact test. No finding was clearly significant in the analysis of both the 144-bp flanking sequences and the 50-bp sequence data. For these reasons, it was important to test some of the hypotheses generated by the sequence analysis by an independent method.

To this end, integration near repeated sequences was studied by using an assay based on PCR amplification of host-virus DNA junctions. In each reaction, one primer was complementary to an HIV-1 LTR end and the second primer was complementary to a repeated sequence (alphoid, *Alu*, LINE-1, or THE 1 repeats) (Fig. 3) (30, 44, 62). The first PCR amplification was followed by a second PCR with nested primers. The LTR primer in the second amplification was labeled at the 5′ end with ^{32}P. Amplification products were separated on DNA sequencing-type gels and analyzed by autoradiography. An integration event in or near the repeated sequence studied gave rise to a labeled band by amplification. Amplification of many such integration events gave rise to a ladder of labeled bands on the final autoradiogram.

The importance of the in vivo setting was assessed by com-

TABLE 2. Integration sites analyzed and their similarities to known sequences

Sequence name[a]	Length (bp)[b]	Dup seq[c]	Identified similarities[d]	Identified similarities truncated to 50 bp[e]
MolH 1	106	ATGTC	*[f]	*
MolH 2	60	CAAGC	*	*
SupH 1	156	TCTTC	LINE-1 [2–153, SW = 508]	*
SupH 2	132	GCTAC	*	*
SupH 3	91	GGAAA	*	*
SupH 4	139	GTGGT	*	*
SupH 5	140	TATAT	*	*
SupH 6	114	ATCCC	*	*
SupH 7	230	GCATG	*	*
SupH 9	82	CTATA	*	*
SupH 10	212	TACAC	LINE-1 [2–107, SW = 251]	*
SupH 11	166	CATGC	*Alu* [15–110, SW = 716]	Alu [SW = 304]
SupH 12	89	GTTGG	*	*
SupH 13	63	CTCAC	Transcription unit (cDNA) [5–62, $P = 1.6 \times 10^{-16}$]	Transcription unit (cDNA) [$P = 1.9 \times 10^{-12}$]
SupH 14	111	GTCAC	*	*
SupH 15	164	TATGG	LINE-1 [2–107, SW = 400]	*
SupH 16	66	AACAG	*	*
SupH 17	54	CTCAC	*	*
SupH 18	159	GTTGT	*	*
SupH 20	342	GTTTC	*Alu* [3–125, SW = 956]	Alu [SW = 373]
SupH 21	173	CATAT	*	*
SupH 22	38	CACAC	*	Excluded
SupH 23	258	CATTC	*	*
SupH 24	110	GTAAT	*	*
SupH 25	37	CTTTT	*	Excluded
SupH 27	160	CCATT	*	*
SupH 28	93	AATAC	Transcription unit (cDNA) [1–93, $P = 3.7 \times 10^{-33}$]	Transcription unit (cDNA) [$P = 1.5 \times 10^{-13}$]
SupH 29	143	GCCCA	*	*
SupH 31	188	ATATT	*	*
SupH 32	157	GTTGA	Transcription unit (cDNA) [59–157, $P = 5.9 \times 10^{-34}$]	*
SupH 33	50	CTTCA	Transcription unit (VACH1 gene) [1–50, $P = 6 \times 10^{-13}$]	Transcription unit (VACH1 gene) [$P = 6 \times 10^{-13}$]
SupH 34	50	AGTTG	*	*
SupH 35	420	TTAAC	Transcription unit (cDNA) [52–143, $P = 2.8 \times 10^{-25}$]; LINE-2 [223–274, SW = 252]	*
SupH 36	237	CTTGT	*	*
SupH 37	69	CACAC	Alu [1–69, SW = 471]	Alu [SW = 371]
SupH 38	68	GTTAT	*	*
SupH 39	89	CAAAA	*	*
SupH 41	41	ATGGC	*	Excluded
SupH 42	437	AAAAC	LINE-1 [1–437, SW = 2684]	LINE-1 [SW = 264]
SupH 43	179	ATAGT	Transcription unit (cDNA) [1–179, $P = 9.4 \times 10^{-65}$]; other repeat (LTR element) [98–152, SW = 198]	Transcription unit (cDNA) [$P = 3.8 \times 10^{-13}$]
SupH 44	337	GAAAC	Other repeat (MIR, SINE) [191–315, SW = 493]	*
SupH 46	81	GGGAG	Transcription unit (cDNA) [1–33, $P = 3.9 \times 10^{-6}$]	Transcription unit (cDNA) [$P = 4.6 \times 10^{-6}$]
SupH 47	111	AAAAC	Transcription unit (cDNA) [1–57, $P = 2.1 \times 10^{-13}$]	Transcription unit (cDNA) [$P = 2.2 \times 10^{-9}$]
SupH 48	125	CTGTG	Other repeat (MIR, SINE) [1–123, SW = 474]	Other repeat (MIR, SINE) [SW = 245]
SupH 49	260	TTTTG	Alu [1–128, SW = 698]	Alu [SW = 300]
SupS 1	176	GCAGG	Transcription unit (CD27 gene) [1–176, $P = 2.7 \times 10^{62}$]	Transcription unit (cDNA) [$P = 5.4 \times 10^{-13}$]
SupS 2	113	GTTCT	*	*
SupS 3	125	ATACC	Alu [4–115, SW = 540]	Alu [SW = 195]
SupS 4	215	CCCTC	Other repeat (MER74, LTR element) [1–213, SW = 599]	Other repeat (MER74, LTR element) [SW = 277]
SupS 5	147	CAGCA	*	*
SupS 7	171	GAGTC	*	*
SupS 8	85	TGAGT	Transcription unit (cDNA) [1–81, 3.2×10^{-26}]	Transcription unit (cDNA) [$P = 3.6 \times 10^{-13}$]
SupS 9	86	GTACC	*	*
SupS 10	52	AAAGC	Alu [2–59, SW = 356]	Alu [SW = 310]
SupS 11	147	CTAAC	*	*
SupS 12	131	GTTTC	*	*
SupS 13	94	ATGTG	Transcription unit (cDNA) [1–94, $P = 5.1 \times 10^{-28}$]	Transcription unit (cDNA) [$P = 3.4 \times 10^{-12}$]
SupS 14	184	GAGAC	*	*
SupS 15	120	AAATG	*	*
SupS 16	161	CTCTG	*	*
SupS 17	215	GTATG	*	*
Total bp	8,809			2,900
Avg	144			50

[a] Laboratory designation for each DNA clone.

[b] Number of human DNA base pairs sequenced adjacent to the HIV cDNA terminus.

[c] Nucleotide sequence of the 5 bp of human DNA at the junction with viral DNA expected to be duplicated upon integration.

[d] Sequence similarities found by comparison to sequence databases (the first designation is the sequence class given in Table 3, the name in parentheses is a more detailed designation, and the numbers in brackets represent the location of the sequence match [e.g., 1 = the first cDNA-proximal base pair in host DNA] and the degree of similarity).

[e] Similarities identified in the 50-bp sequence data set. For explanation of bracketed data, see footnote *d*.

[f] *, anonymous.

TABLE 3. Sequence composition of libraries of integration sites and control DNA fragments

Sequence class	Analysis of 144-bp sequences (avg length)[a]		Reanalysis of 50-bp sequences[b]	
	Integration sites (%)	Genomic DNA (%)	Integration sites (%)	Genomic DNA (%)
Anonymous	61	43	69	71
Alu element	10	9	10	6
LINE element	8	13	2	6
Alphoid repeat	0	6	0	3
Other repeats	7	22	3	10
Transcription unit	18	8	16	4

[a] For data from sequences of 144-bp average length, 61 integration sites and 104 control sequences were considered.

[b] For the reanalysis of integration site sequences considering only the proximal 50 bp of human DNA sequence, 58 integration sites and 104 control sequences were considered.

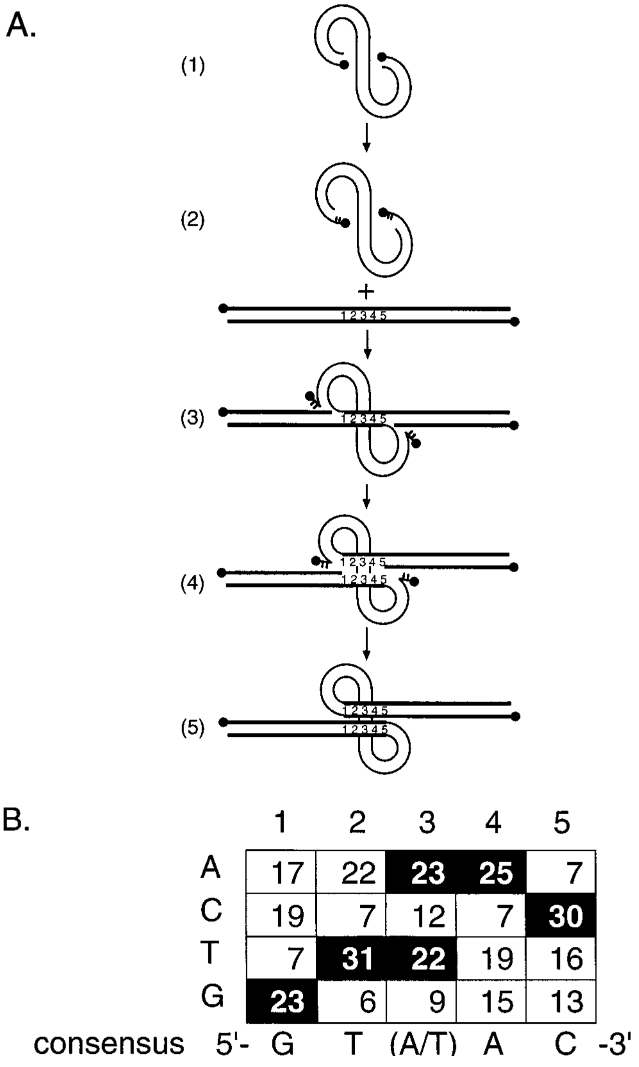
paring integration sites from infected cells with sites made in vitro by integration into deproteinized chromosomal DNA. The in vitro reactions were carried out by using PICs purified from infected cells as a source of integration activity (5, 15, 19). PICs contain the viral cDNA in association with the virus-encoded integrase protein and other viral and cellular proteins (7, 17, 20, 22, 32). Previous studies have demonstrated that incubation of PICs with naked DNA targets results in the covalent integration of some of the HIV cDNA into target (for reviews, see references 13 and 18). The DNA samples from in vivo infections or in vitro integration reactions used for PCR contained similar numbers of proviruses (data not shown).

Amplification of DNA from in vitro integration reactions with the alphoid primer yielded a ladder of labeled bands indicative of integration (Fig. 3B, lane 5). However, amplification of DNA from infected cells with the alphoid primer did not yield a ladder of labeled bands (Fig. 3B, lane 4), indicating that integration did not take place in or near these sequences in vivo. Similar assays using primers complementary to *Alu*1 elements (Fig. 3B, compare lanes 9 and 10), LINE-1 elements (Fig. 3B, compare lanes 14 and 15), and THE 1 repeats (Fig. 3B, compare lanes 19 and 20) yielded integration bands in both in vivo- and in vitro-integrated samples. This finding bolsters the idea that alphoid sequences are competent for integration in naked DNA but masked in vivo. *Alu*, LINE-1, and THE 1 elements, in contrast, are competent in both cases.

Control amplification reactions with no added template DNA (Fig. 3B, lanes 1, 6, 11, and 16) or with DNA from uninfected human T cells did not yield labeled bands (Fig. 3B, lanes 3, 8, 13, and 18). A further control containing integration reactions in vitro carried out in the presence of EDTA to chelate the required metal was mainly negative, although occasional artifactual bands of unknown origin were seen (Fig. 3B, lanes 7 and 12).

Primary DNA sequences favored for integration. Alignment of human DNA sequences at integration junctions yielded a consensus sequence (Fig. 2 and 4). A related sequence has been reported by Stevens and Griffith (55). To determine whether this sequence was favored for integration as naked DNA, several model sequences were synthesized and tested using integration in vitro. Target 1 contained the favored motif embedded in an arbitrary DNA sequence (Fig. 4A, target 1). Target 2 is identical to target 1 except for changes at the two most conserved positions (Fig. 4A, nucleotide positions 1 and 5) from the most favored nucleotide to the least favored. Target 3, like target 1, contained the favored target sequence but embedded in different arbitrary flanking DNA.

Integration assays were carried out to examine favored sites in each sequence. Since previous work indicated that target site selection in naked DNA differed between PICs and the simpler integration complexes formed with recombinant HIV integrase protein (4), the two sources of integration activity were compared. As for the experiment illustrated in Fig. 3, integration products were analyzed by amplification using one primer complementary to the viral DNA end and a second primer



	1	2	3	4	5
A	17	22	**23**	**25**	7
C	19	7	12	7	**30**
T	7	**31**	**22**	19	16
G	**23**	6	9	15	13
consensus 5'-	G	T	(A/T)	A	C -3'

FIG. 2. Consensus sequence at the junctions between HIV cDNA and host DNA and the mechanism of generation of the host sequence duplication. (A) Integration pathway. HIV cDNA is shown as the curved line in part 1. Two nucleotides are removed from each 3′ end of the cDNA (part 2). Host target DNA is shown as a straight line. The host DNA that becomes duplicated is indicated by the numbers 1 to 5. The recessed 3′ ends of the cDNA are then attached to protruding 5′ ends in the target DNA (part 3), and the integration intermediate melts to yield single-stranded gaps at each end (part 4). The in vitro integration reactions with PICs stop at this stage. Repair of the DNA gaps at each host-virus DNA junction results in the production of the 5-bp duplication of target DNA (part 5). (B) Tabulation of the host sequence inferred to be duplicated in our integration site collection. HIV cDNA is joined to target DNA just 5′ of position 1, as illustrated, and similarly on the other strand. Sixty-six duplications are included in this compilation, 61 from the sites listed in Table 2 and 5 additional integration sites with the following duplication sequences: 5′-AGAGT-3′, 5′-GGTAC-3′, 5′-AACAT-3′, 5′-GTAAC-3′, 5′-AATGT-3′ (data not shown).

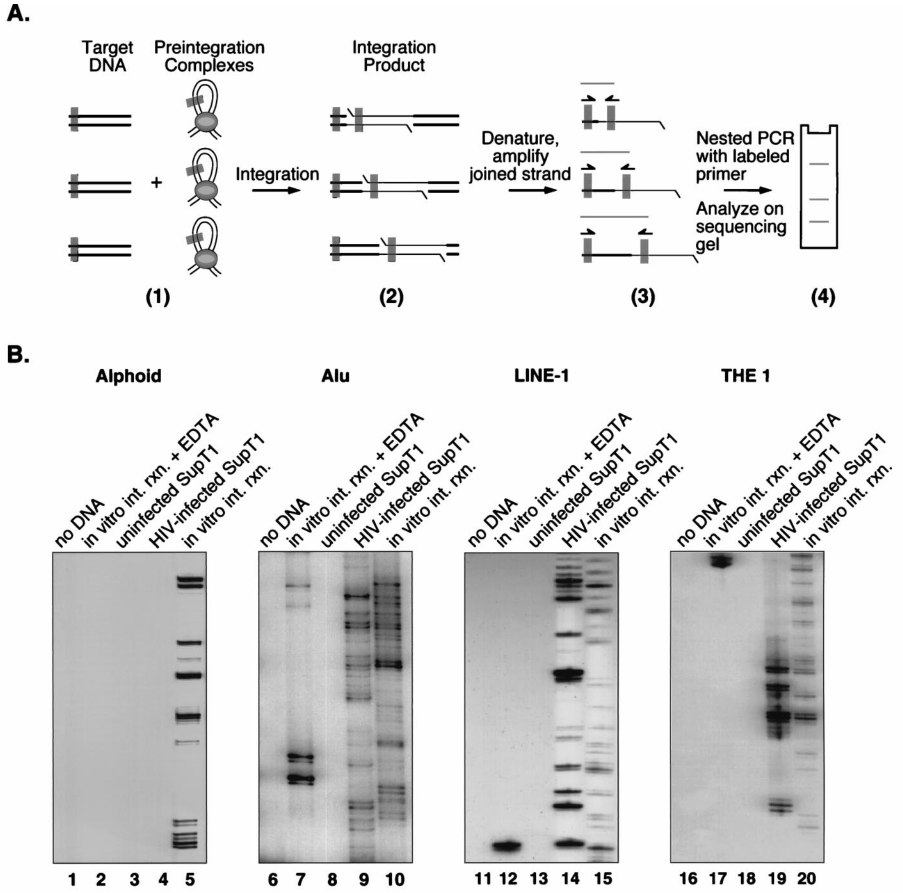


FIG. 3. Analysis of integration sites near several repeat families using a PCR-based assay. (A) Diagram of the PCR method used to analyze integration sites. Primer binding sites are shown as gray rectangles. Part 1 illustrates either integration in vivo into cellular chromosomes or integration in vitro into deproteinized DNA. Products of integration reactions in vitro differ from products made in vivo in that only the former has the DNA breaks indicated in part 2 (the gapped integration intermediate is quickly repaired in vivo). In part 4, the three bands on the sequencing gel arose from three different integration events. (B) Results of PCR assays using primers complementary to alphoid repeats (lanes 1 to 5), *Alu* elements (lanes 6 to 10), LINE-1 elements (lanes 11 to 14), and THE 1 elements (lanes 16 to 20). The presence of a ladder of bands indicates that the template DNA contained HIV cDNA integrated near the repeat family specified. Lanes: 1, 6, 11, and 16, control amplification reactions with no added template; 2, 7, 12, and 17, amplification of inactive PICs and SupT1 DNA; 3, 8, 13, and 18, amplification from uninfected SupT1 DNA; 4, 9, 14, and 19, amplification of DNA from HIV-1 infected SupT1 cells; 5, 10, 15, and 20, amplification of deproteinized DNA that had been incubated with active PICs in vitro. Cellular DNA was detectable as a contaminant of the PIC preparations (data not shown); cellular DNA might have served as an integration target during PIC preparation or participated in recombination during PCR, possibly giving rise to the artifactual bands in lanes 7 and 12.

complementary to target sequences flanking the region of interest. Thus, each band on the final autoradiogram represents integration at a single target phosphodiester, and the intensity of the band represents the relative number of integration events.

Assays of PICs revealed the presence of a strong integration band at the position expected for the hot spot in target 1 (Fig. 4B, lanes 3 and 8). Altering the two most favored bases (target 2) greatly reduced the signal at this position (Fig. 4B, lanes 4 and 9). Assays of target 3, in which the flanking DNA was changed but the favored sequence was preserved, displayed favored integration at the expected hot spot sequence (Fig. 4B, lanes 5 and 10). PCR assays to which no template was added (Fig. 4B, lanes 1 and 6), or which contained mock integration reactions carried out in the presence of EDTA instead of the required divalent metal (Fig. 4B, lanes 2 and 7), revealed no reproducible amplification products. Taken together, these data indicate that the favored target sequence identified from studies in vivo is sufficient to act as a hot spot for PICs in vitro.

Figure 4C presents an analysis of integration directed by purified HIV integrase into targets 1 to 3. The arrows mark the expected location of integration at the hot spot. A band is visible for targets 1 and 3 on the top strand (Fig. 4C, lanes 13 and 15) and bottom strand (Fig. 4C, lanes 18 and 20), although integration by purified integrase at the hot spot for PIC integration is much less prominent. This difference in target site selection highlights the differences between the two sources of integration activity, paralleling previous studies (for review and references, see reference 18).

DISCUSSION

We have used two methods to characterize chromosomal sites used by HIV-1 for integration in human SupT1 cells. We have sequenced a collection of integration sites and a collection of control sites and also analyzed integration near various repetitive sequences by using a PCR-based assay. DNA to be analyzed was prepared only 12 h after initiation of infection in an effort to obtain a population of sites unbiased by subsequent outgrowth of infected cells. In addition, the importance of a conserved host sequence at integration sites was tested by

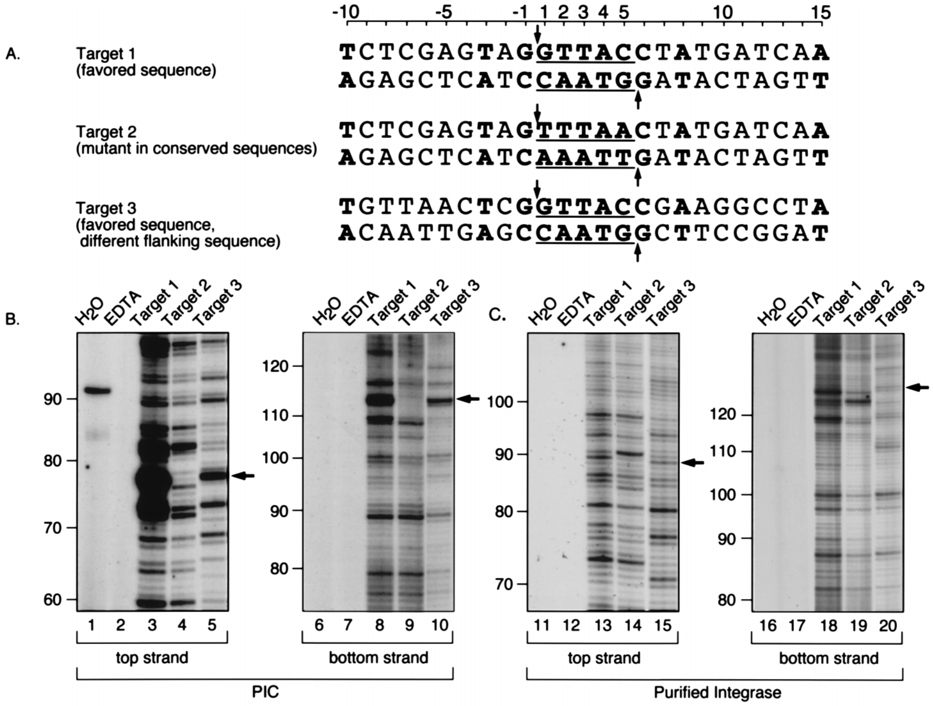


FIG. 4. A conserved sequence at integration sites and analysis of integration at such sites in vitro. (A) Integration target sites tested. The host sequences duplicated upon integration are underlined; the points at which covalent strand transfer takes place on each strand are indicated by arrows; bases favored at integration sites are in boldface type. (B) Integration into targets 1 to 3 directed by PICs. Lanes: 1 and 6, H_2O instead of template; 2 and 7, EDTA added to integration reactions. 3 and 8, target 1; 4 and 9, target 2; 5 and 10, target 3. Arrows indicate the location of the expected integration hotspots (5′ of position 1 on the top strand and 5′ of position 5 on the bottom strand). (C) Integration into targets 1 to 3 directed by purified HIV-1 integrase. Lanes 11 to 20 correspond to lanes 1 to 10, respectively, in panel B. Sizes were assigned by coelectrophoresis adjacent to several DNA sequencing ladders generated by the Sanger method.

using integration in vitro. These studies clarify several factors influencing the selection of chromosomal sites for integration.

Comparison with integration site selection by yeast retrotransposons. Previous studies of Ty retrotransposons in yeast reveal that retroelement integration can be highly site specific. The yeast Ty retrotransposons replicate by transcription, reverse transcription, and integration by using reverse transcriptase and integrase enzymes similar in function and sequence to their retroviral counterparts (2). Ty elements differ from retroviruses in that all steps in replication take place in a single cell. For this reason, Ty retrotransposons must be fastidious in their selection of integration sites, since integration into a required cellular gene would be lethal for the host and suicidal for the transposon.

Ty elements integrate selectively in benign locations in host DNA. Ty1 integrates in a window of several hundred base pairs upstream of host polymerase III (Pol III)-transcribed genes (26). Ty3 is the most selective, integrating at the start site of transcription of Pol III-transcribed genes (12, 29). Ty5 shows a different specificity, integrating in telomeres and in the silent mating cassette DNA (65, 66).

The potential for extreme integration site bias revealed in the Ty studies formed part of the motive for carrying out a large-scale investigation of integration site selection by HIV-1. In humans, integration in Pol III transcription units or telomeric repeats should have been detectable but no strong bias in favor of such sequences was found here or in previous studies with HIV or other retroviruses (23, 45, 47, 55, 58, 62). Evidently, HIV and Ty elements differ in this respect.

Favored integration near active genes? Our data neither strengthen nor exclude the model that integration is favored in open chromatin near active genes (23, 45, 47, 58). Identifiable transcription units were present more frequently in the integration site libraries than in the control libraries. However, the difference was not statistically significant for the 144-bp sequence comparison, although it was significant for the 50-bp sequence comparison (Table 3).

Conclusions concerning integration site location will need to be reevaluated as new information becomes available. It will be particularly interesting to compile and analyze all the known integration site sequences (references 55, 59, and 60 and present study) when the sequence of the human genome is completed and cDNAs and regulatory regions are mapped onto the genomic DNA.

Lack of evidence for favored integration near *Alu* or LINE elements. The data did not indicate that integration was favored near LINE elements or *Alu* elements as previously proposed (54, 55). Both the sequencing study and the region-specific PCR study failed to show any clear biases. One previous proposal was not directly tested. Stevens and Griffith proposed that integration might be favored near *Alu* islands, chromosomal regions containing clustered *Alu* repeats (55). Because our sequencing study examined relatively short flanking sequences (average length, 144 bp), clustering of *Alu* repeats near integration sites could not be assessed.

An effect of primary sequence. The data presented here also reveal a modest favoring of integration at a particular host DNA sequence. Previous studies of integration site sequences have revealed weakly conserved motifs for several retroviruses, including HIV (21, 43, 55). Two mechanisms might account for the observed sequence bias: the integration machinery might interact favorably with a factor bound at the conserved site, or the PIC itself might interact favorably with the conserved sequence as naked DNA. We found that the conserved sequence was favored in vitro as naked DNA, supporting the idea that the conserved sequence is favored in vivo due to interaction with the PIC itself.

Disfavored integration at centromeric alphoid repeats. The most striking feature of our data is the absence of integration in vivo into centromeric alphoid repeats. Alphoid repeats were absent in integration site sequences but present in controls, and alphoid sequences were selectively disfavored in the repeat-specific PCR integration assay. Several lines of evidence indicate that centromeric heterochromatin is organized differently than euchromatin. (i) Heterochromatic centromeres are seen to be more compact than euchromatin in fixed chromosome spreads (6). (ii) Alphoid sequences are more resistant to digestion with DNase I in isolated nuclei than are most DNAs (38, 63). (iii) Alphoid repeats are associated with the centromere-specific proteins CENP-A, CENP-B, and CENP-C (38, 63). On the basis of the data reported here, we propose that HIV-1 cDNA integration is obstructed by packaging DNA in centromeric heterochromatin. These data provide an unexpected demonstration of the long-standing possibility that certain types of chromatin may obstruct cDNA integration.

The mechanism of the integration block is unclear. The wrapping of DNA in heterochromatin may itself provide a steric block to integration, a possibility supported by the observation of condensed structures at centromeres. Other models are also possible. Since gene activity is probably reduced in heterochromatin, HIV may have evolved to avoid integration in heterochromatin to optimize gene expression. Alternatively, centromeric DNA might be sequestered at a nuclear location inaccessible to incoming PICs.

ACKNOWLEDGMENTS

S.C. and C.H. contributed equally to this work.

We thank Gary Karpen, Leslie Orgel, and members of the Bushman laboratory for suggestions and comments on the manuscript, Arian Smit for advice on identifying repeated sequences, and Leslie Barden and Allison Bocksruker for artwork and help in preparing the manuscript.

This work was supported by grants AI 34786 and AI 37489. S.C. was supported in part by the Rau Foundation. F.B. is a Scholar of the Leukemia Society of America.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Boeke, J. D.** 1989. Transposable elements in *Saccharomyces cerevisiae*, p. 335–374. *In* D. E. Berg and M. M. Howe (ed.), Mobile DNA. American Society for Microbiology, Washington, D.C.
3. **Bor, Y.-C., F. Bushman, and L. Orgel.** 1995. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. Proc. Natl. Acad. Sci. USA **92:**10334–10338.
4. **Bor, Y.-C., M. Miller, F. Bushman, and L. Orgel.** 1996. Target sequence preferences of HIV-1 integration complexes in vitro. Virology **222:**238–242.
5. **Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop.** 1987. Correct integration of retroviral DNA in vitro. Cell **49:**347–356.
6. **Brown, S. W.** 1966. Heterochromatin. Science **151:**417–425.
7. **Bukrinsky, M. I., N. Sharova, T. L. McDonald, T. Pushkarskaya, G. W. Tarpley, and M. Stevenson.** 1993. Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. Proc. Natl. Acad. Sci. USA **90:**6125–6129.
8. **Bushman, F., and M. D. Miller.** 1997. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. J. Virol. **71:**458–464.
9. **Bushman, F. D.** 1994. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. Proc. Natl. Acad. Sci. USA **91:**9233–9237.
10. **Bushman, F. D., and R. Craigie.** 1991. Activities of human immunodeficiency virus (HIV) integration protein *in vitro*: specific cleavage and integration of HIV DNA. Proc. Natl. Acad. Sci. USA **88:**1339–1343.
11. **Bushman, F. D., and R. Craigie.** 1992. Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites. Proc. Natl. Acad. Sci. USA **89:**3458–3462.
12. **Chalker, D. L., and S. B. Sandmeyer.** 1992. Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes Dev. **6:**117–128.
13. **Coffin, J. M.** 1996. Retroviridae: the viruses and their replication, p. 1767–1848. *In* B. N. Fields, D. M. Knipe, and R. M. Howley (ed.), Virology. Lippincott-Raven Publishers, Philadelphia, Pa.
14. **Cordonnier, A., J.-F. Casella, and T. Heidmann.** 1995. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related *pol* sequence. J. Virol. **69:**5890–5897.
15. **Ellison, V. H., H. Abrams, T. Roe, J. Lifson, and P. O. Brown.** 1990. Human immunodeficiency virus integration in a cell-free system. J. Virol. **64:**2711–2715.
16. **Fanning, T. G., and M. F. Singer.** 1987. LINE-1: a mammalian transposable element. Biochim. Biophys. Acta **910:**203–212.
17. **Farnet, C., and F. D. Bushman.** 1997. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. Cell **88:**1–20.
18. **Farnet, C. M., and F. D. Bushman.** 1996. HIV cDNA integration: molecular biology and inhibitor development. AIDS **10**(Suppl. A)**:**3–11.
19. **Farnet, C. M., and W. A. Haseltine.** 1990. Integration of human immunodeficiency virus type 1 DNA in vitro. Proc. Natl. Acad. Sci. USA **87:**4164–4168.
20. **Farnet, C. M., and W. A. Haseltine.** 1991. Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. J. Virol. **65:**1910–1915.
21. **Fitzgerald, M. L., and D. P. Grandgenett.** 1994. Retroviral integration: in vitro host site selection by avian integrase. J. Virol. **68:**4314–4321.
22. **Gallay, P., S. Swingler, J. Song, F. Bushman, and D. Trono.** 1995. HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase. Cell **17:**569–576.
23. **Hartung, S., R. Jaenisch, and M. Breindl.** 1986. Retrovirus insertion inactivates mouse a1(I) collagen gene by blocking initiation of transcription. Nature **320:**365–367.
24. **Howard, M. T., and J. D. Griffith.** 1993. A cluster of strong topoisomerase II cleavage sites is located near an integrated human immunodeficiency virus. J. Mol. Biol. **232:**1060–1068.
25. **Hwu, H. R., J. W. Roberts, E. H. Davidson, and R. J. Britten.** 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc. Natl. Acad. Sci. USA **83:**3875–3879.
26. **Ji, H., D. P. Moore, M. A. Blomberg, L. T. Braiterman, D. F. Voytas, G. Natsoulis, and J. D. Boeke.** 1993. Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell **73:**1–20.
27. **Kass, D., M. Batzer, and P. Deininger.** 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol. Cell. Biol. **15:**19–25.
28. **Kimpton, J., and M. Emerman.** 1992. Detection of replication-competent and pseudotyped human immunodeficiency virus with a sensitive cell line on the basis of activation of an integrated β-galactosidase gene. J. Virol. **66:**2232–2239.
29. **Kirchner, J., C. M. Connolly, and S. B. Sandmeyer.** 1995. In vitro position-specific integration of a retroviruslike element requires Pol III transcription factors. Science **267:**1488–1491.
30. **Kitamura, Y., Y. M. Lee, and J. M. Coffin.** 1992. Nonrandom integration of retroviral DNA in vitro: effect of CpG methylation. Proc. Natl. Acad. Sci. USA **89:**5532–5536.
31. **Lewis, P., M. Hensel, and M. Emerman.** 1992. Human immunodeficiency virus infection of cells arrested in the cell cycle. EMBO J. **11:**3053–3058.
32. **Miller, M. D., C. M. Farnet, and F. D. Bushman.** 1997. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J. Virol. **71:**5382–5390.
33. **Miller, M. D., B. Wang, and F. D. Bushman.** 1995. Human immunodeficiency virus type 1 preintegration complexes containing discontinuous plus strands are competent to integrate in vitro. J. Virol. **69:**3938–3944.
34. **Milot, E., A. Belmaaza, E. Rassart, and P. Chartrand.** 1994. Association of a host DNA structure with retroviral integration sites in chromosomal DNA. Virology **201:**408–412.
35. **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.
36. **Paulson, K. E., N. Deka, C. W. Schmid, and L. Leinwand.** 1985. A transposon-like element in human DNA. Nature **316:**359–361.
37. **Pauza, C. D.** 1990. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. Virology **179:**886–889.
38. **Pluta, A. R., A. M. Mackay, A. M. Ainsztein, I. G. Goldberg, and W. C. Earnshaw.** 1995. The centromere: hub of chromosomal activities. Science **270:**1591–1594.
39. **Pognan, F., and C. Paoletti.** 1990. A new extraction procedure of autono-

mous DNA from eucaryotic cells, where DNA could be bound to proteins. Nucleic Acids Res. **18:**5571–5572.

40. **Pruss, D., F. D. Bushman, and A. P. Wolffe.** 1994. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. Proc. Natl. Acad. Sci. USA **91:**5913–5917.
41. **Pruss, D., R. Reeves, F. D. Bushman, and A. P. Wolffe.** 1994. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J. Biol. Chem. **269:**25031–25041.
42. **Pryciak, P., H.-P. Muller, and H. E. Varmus.** 1992. Simian virus 40 minichromosomes as targets for retroviral integration in vivo. Proc. Natl. Acad. Sci. USA **89:**9237–9241.
43. **Pryciak, P. M., A. Sil, and H. E. Varmus.** 1992. Retroviral integration into minichromosomes in vitro. EMBO J. **11:**291–303.
44. **Pryciak, P. M., and H. E. Varmus.** 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell **69:**769–780.
45. **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.
46. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.
47. **Scherdin, U., K. Rhodes, and M. Breindl.** 1990. Transcriptionally active genome regions are preferred targets for retrovirus integration. J. Virol. **64:**907–912.
48. **Scottoline, B. P., S. Chow, V. Ellison, and P. O. Brown.** 1997. Disruption of the terminal base pairs of retroviral DNA during integration. Genes Dev. **11:**371–382.
49. **Sels, F. T., S. Langer, A. S. Schulz, J. Silver, M. Sitbon, and R. W. Friedrich.** 1992. Friend murine leukaemia virus is integrated at a common site in most primary spleen tumours of erythroleukaemic animals. Oncogene **7:**643–652.
50. **Shih, C.-C., J. P. Stoye, and J. M. Coffin.** 1988. Highly preferred targets for retrovirus integration. Cell **53:**531–537.
51. **Siebert, P. D., A. Chenchik, D. E. Kellog, K. A. Lukyanov, and S. A. Lukyanov.** 1995. An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. **23:**1087–1088.
52. **Smit, A. F. A.** 1993. Identification of a new, abundant superfamily of mammalian LTR retrotransposons. Nucleic Acids Res. **21:**1863–1872.
53. **Smit, A. F. A.** 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. **6:**743–748.
54. **Stevens, S. W., and J. D. Griffith.** 1994. Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. Proc. Natl. Acad. Sci. USA **91:**5557–5561.
55. **Stevens, S. W., and J. D. Griffith.** 1996. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J. Virol. **70:**6459–6462.
56. **Swingler, S., P. Gallay, D. Camaur, J. Song, A. Abo, and D. Trono.** 1997. The Nef protein of human immunodeficiency virus type 1 enhances serine phosphorylation of the viral matrix. J. Virol. **71:**4372–4377.
57. **Varmus, H. E., and P. O. Brown.** 1989. Retroviruses, p. 53–108. *In* D. E. Berg and M. M. Howe (ed.), Mobile DNA. American Society for Microbiology, Washington, D.C.
58. **Vijaya, S., D. L. Steffan, and H. L. Robinson.** 1986. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. J. Virol. **60:**683–692.
59. **Vincent, K. A., D. York-Higgins, M. Quiroga, and P. O. Brown.** 1990. Host sequences flanking the HIV provirus. Nucleic Acids Res. **18:**6045–6047.
60. **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.
61. **Waye, J. S., and H. F. Willard.** 1985. Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human chromosome. Nucleic Acids Res. **13:**2731–2743.
62. **Withers-Ward, E. S., Y. Kitamura, J. P. Barnes, and J. M. Coffin.** 1994. Distribution of targets for avian retrovirus DNA integration in vivo. Genes Dev. **8:**1473–1487.
63. **Wolffe, A. P.** 1995. Histone deviants. Curr. Biol. **5:**452–454.
64. **Zhang, J. W., W. F. Song, Y. J. Zhao, G. Y. Wu, and G. Stamatoyannopoulos.** 1993. Molecular characterization of a novel form of (A gamma delta beta) zer thalassemia deletion in a Chinese family. Blood **81:**1624–1629.
65. **Zou, S., and D. F. Voytas.** 1997. Silent chromatin determines target preferences of the Saccharomyces retrotransposon Ty5. Proc. Natl. Acad. Sci. USA **94:**7412–7416.
66. **Zou, S., D. A. Wright, and D. F. Voytas.** 1995. The Saccharomyces Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus HMR. Proc. Natl. Acad. Sci. USA **92:**920–924.