Journal of Virology, May 1998, p. 4005–4014 Vol. 72, No. 5
0022-538X/98/$04.00+0

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

Oligo-nucleotide	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

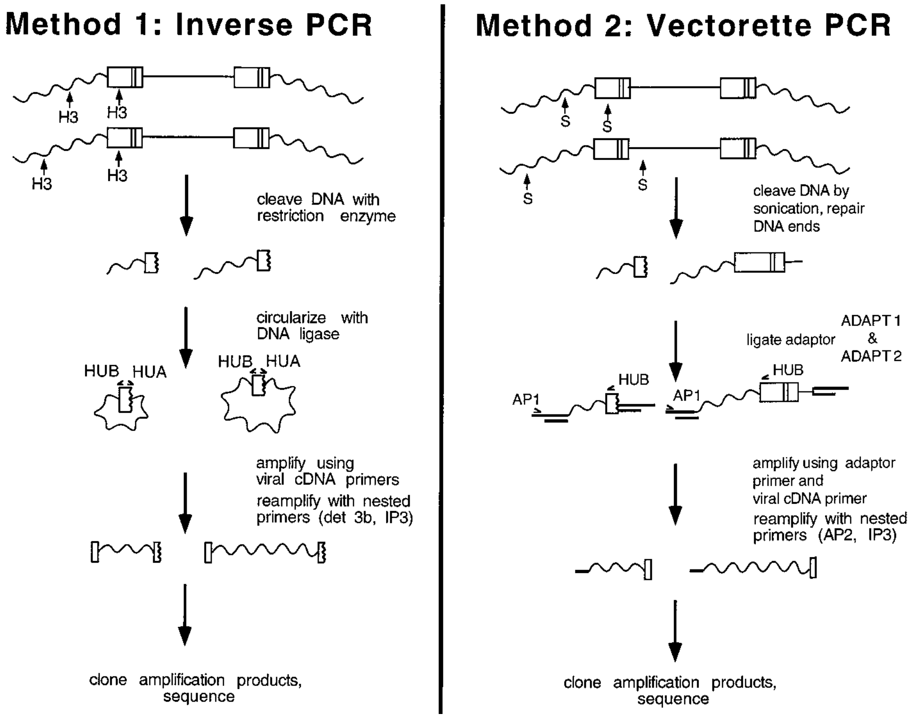
Method 1: Inverse PCR
H3
H3
H3
H3
cleave DNA with restriction enzyme
circularize with DNA ligase
HUB HUA
HUB HUA
amplify using viral cDNA primers
reamplify with nested primers (det 3b, IP3)
clone amplification products, sequence
Method 2: Vectorette PCR
S
S
S
S
cleave DNA by sonication, repair DNA ends
ligate adaptor
ADAPT 1 & ADAPT 2
HUB
HUB
AP1
AP1
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 $[\gamma\text{-}^{32}P]$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.

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**