JOURNAL OF VIROLOGY, Nov. 2006, p. 11313–11321 Vol. 80, No. 22
0022-538X/06/$08.00+0 doi:10.1128/JVI.01737-06

A High-Throughput Method for Cloning and Sequencing Human Immunodeficiency Virus Type 1 Integration Sites▽

Sanggu Kim,[1] Yein Kim,[2] Teresa Liang,[2] Janet S. Sinsheimer,[3] and Samson A. Chow[1,2]*

Biomedical Engineering Interdepartmental Degree Program,[1] Departments of Human Genetics and Biomathematics,[3] and Department of Molecular and Medical Pharmacology, Molecular Biology Institute, and UCLA AIDS Institute,[2] UCLA School of Medicine, Los Angeles, California 90095

Received 10 August 2006/Accepted 31 August 2006

Integration of retroviral DNA is nonspecific and can occur at many sites throughout chromosomes. However, the process is not uniformly distributed, and both hot and cold spots for integration exist. The mechanism that determines target site specificity is not well understood. Because of the nonspecific and widespread nature of integration, studies analyzing the mechanism and factors that control target site selection require the collection and analysis of a large library of human immunodeficiency virus type 1 (HIV-1) proviral clones. Such analyses are time-consuming and labor-intensive using conventional means. We have developed an efficient and high-throughput method of sequencing and mapping a large number of independent integration sites in the absence of any selection or bias. The new assay involves the use of a modified HIV-1 (NL-Mme) containing a type IIS restriction site, MmeI, at the right end of viral DNA. Digestion of genomic DNA from NL-Mme-infected cells generated viral DNA-containing fragments of a discrete size. Subsequent ligation-mediated PCR yielded short integration site fragments termed Int-tags, which were concatemerized for determining multiple integration sites in a single sequencing reaction. Analysis of chromosomal features and sequence preference associated with integration events confirmed the validity of the new high-throughput assay. The assay will aid the effort in understanding the mechanisms of target site selection during HIV-1 DNA integration, and the described methodology can be adapted easily to integration site studies involving other retroviruses and transposons.

Integration of retroviral DNA is nonspecific but not completely random (for reviews, see references 4, 5, 7, and 19), and the frequencies of use for specific sites can vary considerably (31, 37, 42). The site of integration has significant implications for both the virus and the host cell. For virus replication, human immunodeficiency virus type 1 (HIV-1) gene expression is determined by the interplay between cellular transcription factors and the virally encoded Tat, as well as by mRNA splicing and transport (13, 22, 41). In addition, depending on the site of integration, proviruses derived from an identical molecular HIV-1 clone can cause a drastic difference in level of gene expression and thus ability to produce progeny (20, 21, 27). Chromatin of the host genome represents an extremely heterogeneous environment, both functionally and structurally (23). Conceivably, the expression status of a particular provirus may also change depending on chromatin remodeling, going from active to latent or vice versa.

In the host cell, since integration of retroviral DNA is inherently a mutagenic event, a better understanding of the process of target site selection may also provide a better assessment of cellular toxicity induced by insertional mutagenesis (3, 29). The information may also be used to optimize retrovirus-based vectors in genetic engineering and therapy. For instance, vectors with an integration site preference for intergenic region may be more attractive for human gene therapy than those favoring integration near transcription start sites or in active genes (12, 30, 31, 42).

The mechanism that determines target site specificity of retroviruses is not well understood and is likely affected by multiple factors. These factors include the virus-encoded enzyme integrase (IN) that catalyzes the integration reaction, target DNA sequence and structure, transcriptional status of DNA, DNA methylation, repetitive elements, and DNA-binding proteins (see references 6, 7, 11, and 19 and references therein). Because of the nonspecific nature of integration, studies on understanding the mechanism and characterizing the factors that control target site selection in infected cells require the collection and analysis of a large library of HIV-1 proviral clones. Although the roles of the aforementioned factors in the selection of DNA sites for integration have been studied to various extents in vitro, characterization of the roles of many of these factors in vivo is either scarce or inadequate or has not been done (6, 19). The conventional means for analyzing integration sites involve sequencing and mapping of one positive clone for each integration site (for examples, see references 17, 26, 31, 37, and 42). As such, integration site analysis is typically time-consuming and labor-intensive, and owing to the relatively small number of integration events that can be analyzed by these methods, associating integration sites with particular genomic regions or individual genes and evaluating the role of any one factor in integration are not trivial tasks.

We have developed and validated a high-throughput, efficient, and unbiased method of sequencing and mapping a large number of independent integration sites. This high-throughput

* Corresponding author. Mailing address: Department of Molecular and Medical Pharmacology, Molecular Biology Institute, and UCLA AIDS Institute, UCLA School of Medicine, Los Angeles, CA 90095. Phone: (310) 825-9600. Fax: (310) 825-6267. E-mail: schow@mednet.ucla.edu.

▽ Published ahead of print on 13 September 2006.

assay should aid the effort in understanding the mechanisms of target site selection and examining the role of viral factors and host cell processes in influencing the choice of target sites during retroviral DNA integration in vivo.

MATERIALS AND METHODS

Cells and reagents. 293T cells and CEM cells were obtained from the American Type Culture Collection, grown in Dulbecco's modified Eagle's medium (Gibco-BRL) and RPMI 1640, respectively, and supplemented with 10% fetal bovine serum (Omega Sci.), 100 U/ml of penicillin, and 0.1 mg/ml of streptomycin. High-performance liquid chromatography-purified oligonucleotides were purchased from Integrated DNA Technology, *Pfu*Ultra and Herculase DNA polymerases from Stratagene, T4 DNA ligase and restriction enzymes from New England BioLabs and Invitrogen, and Dynabeads from Dynal Biotech (Norway).

Preparation of WT and MmeI-containing viruses. The mutant virus NL-Mme was derived from the wild-type (WT) HIV-1 molecular clone NL4-3 by the introduction of point mutations at positions 630 (T to C) and 632 (G to A) at the 3′ end of the left long terminal repeat (LTR) (Fig. 1A) to create the MmeI recognition site. The primers used for mutagenesis are MmeF (5′-CAGTGTG GAAAATCTCCAACAGTGGC) and MmeR (5′-TGTTCGGGCGCCACTGT TGGAGATTT). pNL4-3 or pNL-Mme DNA was transformed into TOP10 cells (Invitrogen), which were cultured at 30°C, and plasmid DNAs were prepared using an Endofree plasmid maxi kit (QIAGEN). All viral stocks were prepared by PolyFect (QIAGEN) transfection of 1×10^6 293T cells with 4 μg of DNA in 25-cm^2 flasks (1). Culture supernatants were collected 48 h after transfection and passed by gravity through a 0.45-μm low-protein-binding membrane (Corning). Virions were treated with 200 U of RNase-free DNase I (Amersham Pharmacia) per ml of viral stock in the presence of 10 mM $MgCl_2$ at room temperature for 1 h and stored at −80°C until use. The virus titer was estimated by an enzyme-linked immunosorbent assay (Coulter, Inc.) against the HIV-1 p24 antigen.

Assays for cloning and sequencing HIV-1 integration sites. Fifty million CEM cells in 175-cm^2 flasks were infected with either WT or NL-Mme viruses with a multiplicity of infection (MOI) of 10. After 8 h, the culture supernatant containing free viruses was removed and replaced with fresh media. The genomic DNA of infected cells was isolated 48 h postinfection by use of 1% sodium dodecyl sulfate and 100 μg/ml proteinase K and extracted with phenol-chloroform (1). To remove circular (one- and two-LTR circles) and linear, unintegrated viral DNA, 1 to 2 mg genomic DNA was electrophoresed on a 0.4% agarose gel at 4°C. DNA larger than 10 kbp was extracted, purified, and then digested with BamHI and XhoI. BamHI and XhoI cleave once in the viral genome at nucleotide (nt) positions 8466 and 8888, respectively, and were used to produce DNA fragments containing the right LTR and the neighboring cellular DNA. After digestion, the DNA was denatured and annealed with a biotinylated (b) primer, bNLR (5′-b-GTGCCTGGCTAGAAGCACAAG), which is complementary to nt positions 8950 to 8970 within the *nef* gene, about 100 bp upstream of the right LTR. The annealed primer was then extended using *Pfu*Ultra DNA polymerase and deoxynucleoside triphosphates (dNTPs). After chain elongation, the biotin-labeled DNA, which is termed Int-DNA and contains the viral and cellular DNA sequences at the integration site, was isolated from the sample by binding to streptavidin-agarose Dynabeads. The Int-DNA was then processed by two different methods. For simplicity, one method was termed the "conventional" assay and is similar to previously described integration site assays that yield one integration site per positive clone (26, 37, 42). The other method was the high-throughput Int-tag assay described in this report, and it yields multiple integration sites per positive clone.

(i) Conventional assay. The streptavidin-bound Int-DNA was digested with NspI (PuCATG ↓ Py), a 6-bp cutter that produces DNA fragments on average of about 1 kbp in length (Fig. 1B). The digested products were ligated with a short DNA linker (NP linker), which was prepared by annealing BHLinkA (5′-CGG ATCCCGCATCATATCTCCAGGTGTG) with NPLink (5′-CACCTGGAGAT ATGATGCGGGATCCG**CATG**). The NP linker contains a BamHI site (underlined) and a 4-nt 3′-overhang (in bold type) complementary with the NspI-digested Int-DNA fragments. The ligated products were amplified by PCR using U51 (5′-TGGCTAACTAGGGAACCCACT) and NP1 (5′-TCACACCTGGAG ATATGATGCG) as the forward and reverse primers, respectively. U51 anneals to nt positions 9570 to 9590 of the viral U5 end, whereas NP1 anneals to the NP linker. The PCR amplification was carried out in a final volume of 200 μl with 0.5 μM of each primer, 0.2 mM of dNTPs, $\sim 10^{-3}$ pmol of template DNA, and 10 U Herculase DNA polymerase under the following conditions: 2 min of preincubation at 94°C, followed by 27 cycles at 94°C for 30 s, 58°C for 30 s, and 72°C for 5 min. The reaction mixture was then incubated for 10 min at 72°C for a final extension. The PCR products were separated on a 2% agarose gel. DNAs over 100 bp in length were isolated using a gel extraction kit (QIAGEN) and cloned using a Zero Blunt TOPO PCR kit (Invitrogen).

(ii) High-throughput Int-tag assay. The streptavidin-bound Int-DNA was digested by MmeI and then ligated with a 28-bp DNA linker (BH linker) containing a randomized 2-nt 3′-overhang and a 3-nt 5′-overhang in the minus strand (Fig. 1B). The 2 nt at the 3′-overhang were randomized so that the linker could base pair and ligate with the MmeI-digested Int-DNA fragments. The BH linker was prepared by annealing BHLinkA with BHLinkB (5′-TGTCACACCTGGA GATATGATGCGGGATCCGNN). BHLinkB was synthesized to contain a uniform distribution of the 16 possible combinations for two sequential random nucleotides. The random nature of the 2 nucleotides at the 3′ end was confirmed by electrospray ionization mass spectrometry (data not shown). The BH linker contains a BamHI site (underlined) and is not phosphorylated to avoid self-ligation.

The linker-ligated DNA was amplified by a two-step PCR using the forward primer BMF (5′-TCAGACGGATCCAGTCAGTGTGGAAAATCTCC) and the reverse primer NP1. BMF contains a BamHI site (underlined) and anneals to 4 nt upstream of the viral U5 end. Approximately 10^{-4} pmol of template DNA was added to a final volume of 1 ml of 1× *Pfu*Ultra buffer containing 1 μM of BMF and NP1 primers, 0.2 mM of dNTPs, and 20 U *Pfu*Ultra DNA polymerase. The reaction mixture was aliquoted into a 96-well plate for the amplification. The first-round-PCR condition was 2 min of preincubation at 94°C, followed by 27 cycles at 94°C for 30 s, 58°C for 30 s, and 72°C for 1 min, and then a final extension of 10 min at 72°C. After PCR, the reaction mixture was concentrated by ethanol precipitation and separated on a 14% native polyacrylamide gel. The 79- and 80-bp products were extracted and subjected to 20 additional cycles of linear amplification under the condition of the first-round PCR. The amplified DNA was digested with BamHI to form 46- or 47-bp fragments (termed Int-tags) with a 4-nt 5′-overhang at each end (Fig. 1B). Each Int-tag has 25 bp of viral end DNA, 19 or 20 bp of cellular DNA, and 2 bp of linker sequence. The digestion mixture was concentrated and separated on a 14% native polyacrylamide gel at 4°C and 150 V to prevent denaturation of Int-tags. The Int-tags were excised and extracted from the gel and concatemerized by ligation using 10,000 U of T4 DNA ligase (New England BioLabs) in a total of 1 ml reaction mixture incubated at 16°C for 4 h. The reaction mixture was then concentrated to 20 μl and separated on a 12% native polyacrylamide gel. DNA products longer than 400 bp were isolated. Concatemerized Int-tags were cloned into circularized pCR4Blunt previously cut with BamHI. Insertion of concatemerized Int-tags into the BamHI site within pCR4Blunt would disrupt the toxin-producing *ccdB* gene (14) and thus provide positive selection for clones containing Int-tag concatemers.

Sequence analysis and mapping integration sites. The sequence of the cloned DNA was determined by dideoxy sequencing, and sequencing ambiguities were resolved by repeated sequencing on both strands. Cellular DNA sequences obtained from each authentic integration site were cataloged using the software program MacVector 7.1.1 (Oxford Molecular). The sequences were aligned and searched for consensus sequence and features by use of AssemblyLIGN, with the threshold parameter set at 40%.

The chromosomal location of the integration site sequence was mapped to the human genome (Human May 2004 [hg17] assembly, National Center for Biotechnology Information [NCBI] Build 35) by use of the BLASTN program (http://www.ensembl.org) or BLAT (University of California, Santa Cruz; http://genome.ucsc.edu). Transcription units in the vicinity of the integration sites were identified using the RefSeq gene database (NCBI Reference Sequence Project; www.ncbi.nih.gov/RefSeq/). Similarities to repetitive sequences were ranked using the Smith-Waterman parameter generated by Repeat Masker (http://www.repeatmasker.org/).

Statistical analysis of integration site sequences. All statistical analyses were conducted using Stata statistical software (Stata Corp., College Station, TX; www.stata.com). To test for differences in proportions, we used rxc contingency table analysis (by Fisher's exact test when individual cell counts were small [<10] or by chi-square approximation). To test for equality of distribution, we used the two-sample Kolmogorov-Smirnov test.

Nucleotide sequence accession numbers. The GenBank accession numbers for integration sites sequenced in this study are EF035624 through EF035928 for HIV WT and EF035929 through EF036245 for NL-Mme. The integration site sequences shorter than 50 bp discussed in this paper have been deposited on our laboratory website (http://labs.pharmacology.ucla.edu/chowlab/Web/).

RESULTS

Replication kinetics and integration site preference of NL-Mme. A key feature of the new high-throughput assay is the

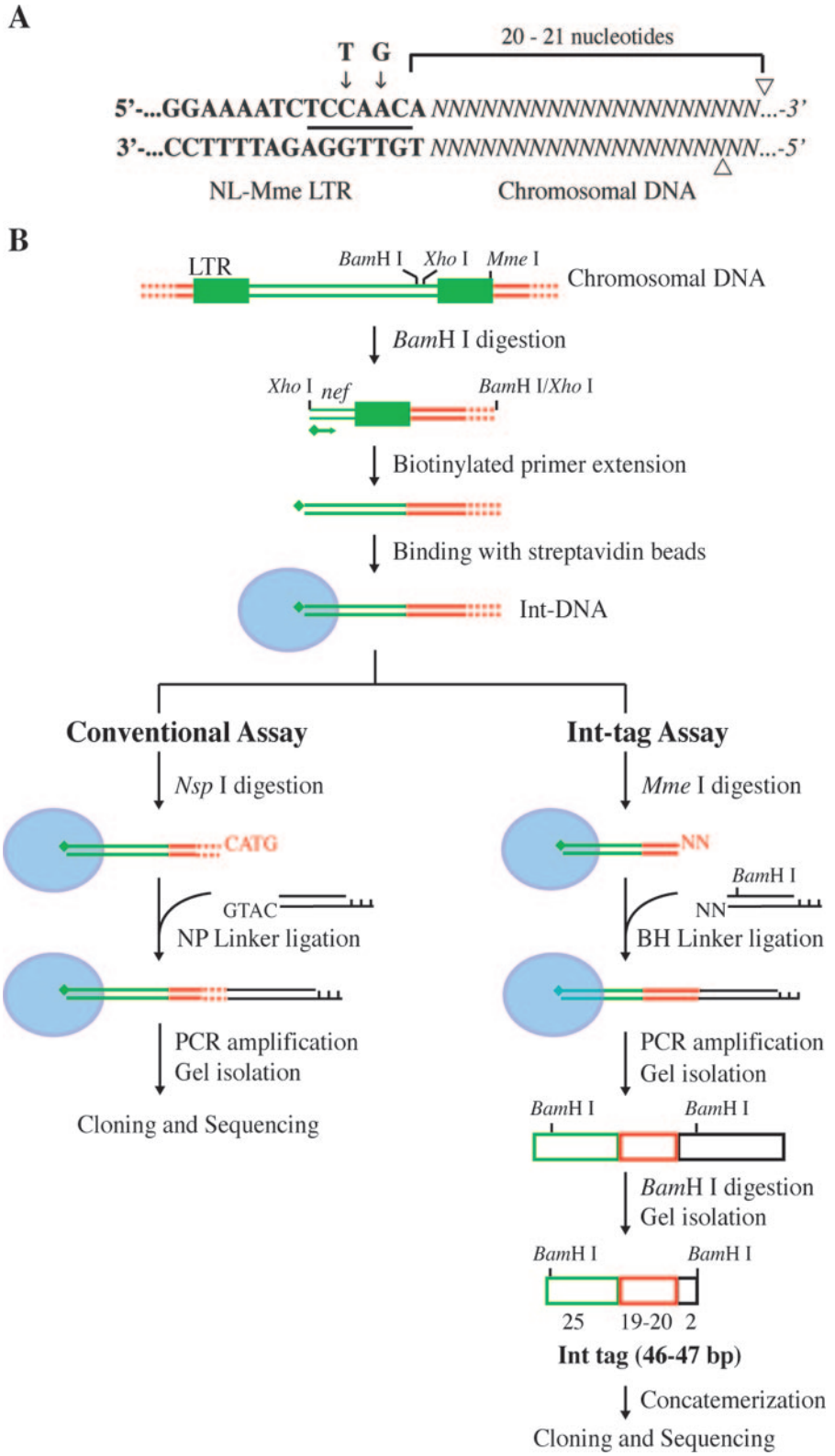

FIG. 1. Assays for genome-wide analysis of HIV-1 integration sites. (A) Construction of mutant HIV with a type IIS MmeI restriction site in the LTR. Bold letters denote viral DNA sequences at the U5 region of the LTR, and italicized letters denote chromosomal DNA. The nucleotides at the positions indicated by the arrows were changed from T and G in the wild-type sequence to C and A, respectively, in the NL-Mme mutant, generating a new recognition site for MmeI (underlined). Arrowheads indicate cleavage sites for MmeI. (B) Schematic diagram outlining the major steps of the conventional assay and the high-throughput Int-tag assay. Viral, cellular, and linker DNAs are denoted by green, red, and black lines or boxes, respectively. Red dotted lines denote cellular DNA with various lengths. Blue ovals represent streptavidin beads, and green diamonds represent biotin. See Materials and Methods for a detailed description of the experimental procedures.

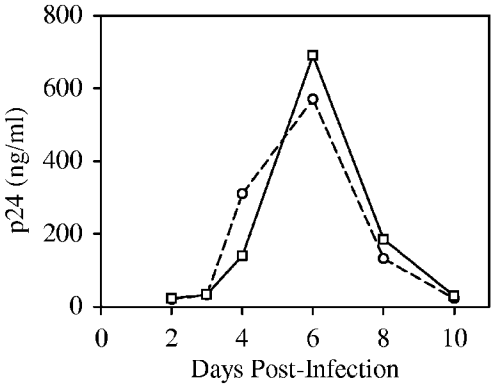

FIG. 2. Replication kinetics of WT and NL-Mme viruses. CEM cells were infected with equal amounts of the p24 equivalent of WT (○) or NL-Mme (□) virus at an MOI of 0.001. The culture media were monitored for p24 levels (ng/ml) at the indicated time points postinfection.

introduction of a type IIS restriction site, MmeI, at the U5 end of the left LTR by replacement of nucleotides T and G at positions 630 and 632 with C and A, respectively (Fig. 1A). Reverse transcription and integration of the MmeI-containing HIV-1 (NL-Mme) placed the MmeI recognition sequence 1 bp away from the right end of the integrated provirus (nucleotide position 9708). MmeI makes a 2-nucleotide 3′-staggered cut 20 or 21 bp downstream from its nonpalindromic recognition sequence, 5′-TCCRAC-3′ (Fig. 1A). To ensure that introduction of the MmeI recognition site in the U5 end of HIV-1 did not affect viral replication, WT or NL-Mme viruses were used to infect CEM cells at an MOI of 0.001. The replication kinetics of NL-Mme, as measured by the p24 level over a period of 10 days, was similar to that of the WT (Fig. 2). The result is consistent with previous findings that base substitution at position 630 is well-tolerated (2, 9, 28, 33, 38) and that mutation of the G at position 632 to A has little effect on the catalytic activity of HIV-1 IN (24, 38). The virus from the culture medium 6 days postinfection was also collected, and the U5 region of the left LTR was sequenced to confirm the retention of the MmeI site (data not shown).

In addition to replication kinetics, we analyzed the distribution of and chromosomal features associated with integration events of the WT and NL-Mme viruses in the human genome. To generate DNA samples containing integrated proviruses, we used the WT or the NL-Mme virus to infect CEM cells at an MOI of 10 and isolated the cellular DNA 2 days after infection. The 2-day period was chosen to minimize clonal expansion and selection but allow sufficient time for infection and integration. Genomic DNA from infected cells was cleaved with restriction enzymes, the virus-host DNA junctions were amplified by a linker-mediated PCR method that we termed the "conventional" assay (Fig. 1B), and the amplified product was cloned and sequenced. This method, similarly to those described previously, yielded one integration site per positive clone (26, 37, 42). The following criteria were used to verify the authenticity of the integration site sequence: (i) the sequence contained both right LTR and linker sequence, (ii) a match to the human genome started after the end of the right LTR (5′-…CA-3′) and ended with the linker sequence, and (iii) the host DNA region from the putative integration site sequence showed 98% or greater identity to the human genomic sequence.

A total of 309 and 323 integration sites from cells infected with the WT and NL-Mme viruses, respectively, were analyzed and mapped to the human genome by use of BLAT. For both WT and NL-Mme viruses, integration events were found in all 23 human chromosomes (22 autosomes and the sex chromosome X) (Fig. 3). Similarly to previously published reports (26, 37), the frequencies of integration of WT HIV-1 were quite different among the different chromosomes in comparison to uniformly random integration ($P = 7.07 \times 10^{-10}$). Notably, chromosomes 12, 17, and 19 were significantly overrepresented (P values of 0.0410, 0.0008, and <0.0001, respectively), while chromosomes 8 and X were significantly underrepresented (P values of 0.0126 and 0.0020, respectively). The frequencies of integration events of NL-Mme in different chromosomes were not different from those of the WT ($P = 0.282$) (Fig. 3).

HIV-1 integration favors transcription units and the *Alu* repeats of the short interspersed nuclear element (SINE) family and disfavors L1 of the long interspersed nuclear element (LINE) and LTR elements (LTR-E) (37, 42). Under our described conditions using CEM cells infected with the WT NL4-3 virus (Fig. 4), we reached the same conclusion for integration pattern comparison ($P = 0.043$) and for the comparison of the proportions of transcription units ($P = <0.0001$). The chromosome features associated with NL-Mme integra-

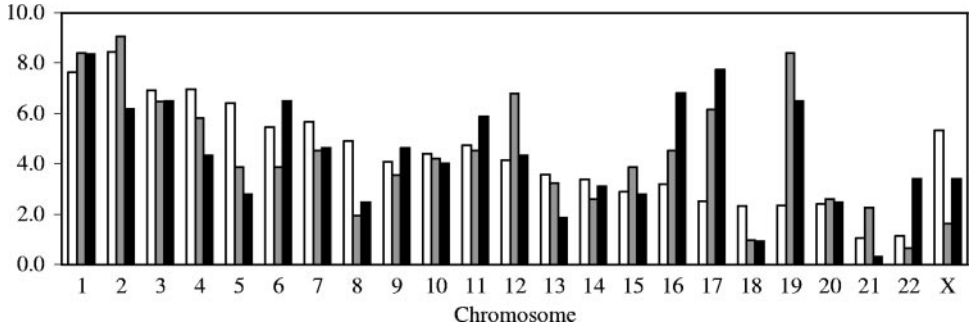

FIG. 3. Distribution of WT and NL-Mme virus integration events in human chromosomes. Results are expressed as the percentages of integration events in each chromosome. Human chromosome numbers are indicated at the bottom of the figure. The numbers of integration events for the random control (open bars), WT virus (gray bars), and NL-Mme virus (black bars) were 5,000, 309, and 323, respectively.

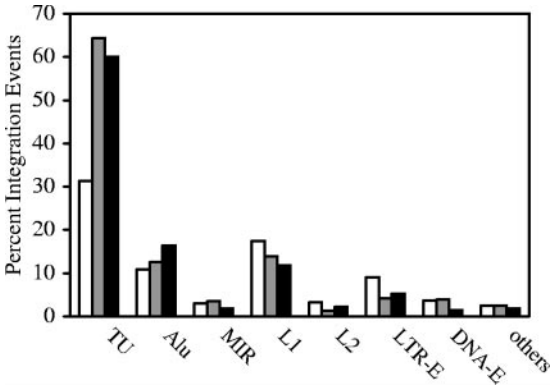

FIG. 4. Analysis by the conventional assay of chromosomal features associated with WT and NL-Mme virus integration events. CEM cells were infected with WT or NL-Mme virus at an MOI of 10. Integration sites were mapped using the conventional assay, and chromosomal features associated with WT (gray bars) and NL-Mme (black bars) proviruses were analyzed. The results are expressed as percentages of total integration events and compared with those of the random control (open bars). Chromosomal features analyzed include transcription units (TU), *Alu* and mammalian interspersed repeat (MIR) of the SINE, L1 and L2 of the LINE, LTR-E, and DNA-E.

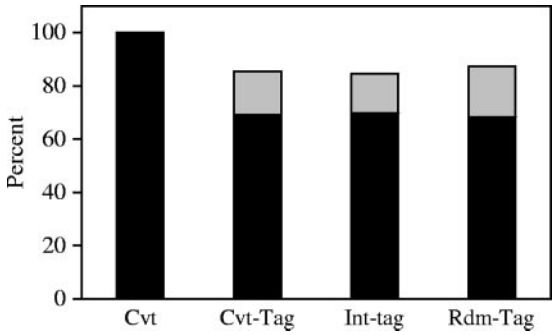

FIG. 5. Mapping of short integration sequences to unique locations and those in identifiable repeat elements. Integration site sequences from cells infected with NL-Mme were determined using the conventional assay (Cvt) or the high-throughput assay (Int-tag). Each sequence was mapped to either unique locations (black bars) or identifiable repeat elements (gray bars), and the results are expressed as percentages of total integration site sequences. The results were compared with those obtained from conventional tags (Cvt-tag), generated by taking the first 19 or 20 nucleotides immediately adjacent to the viral DNA of each integration site sequence determined by the conventional assay, or a random library of 19- and 20-bp cellular sequences (Rdm-tag) generated in silico.

tion events were also similar to those of the WT (Fig. 4), with a *P* value of 0.176 for integration pattern comparison and a *P* value of 0.286 for the comparison of the proportions of transcription units. Based on the replication kinetics and integration preference, we therefore conclude that NL-Mme behaves identically to the WT virus and that introducing the MmeI restriction site to the U5 region of the LTR has no discernible effect.

Sequencing analysis and mapping of integration sites by use of the high-throughput Int-tag assay. To test the new high-throughput method, the identical DNA sample isolated from CEM cells infected with NL-Mme was analyzed using the Int-tag assay and the results were compared to those obtained using the conventional assay. For the Int-tag assay, an added criterion for verifying the authenticity of the integration site sequence was that the length of the intervening sequence between the viral and linker sequences was 19 or 20 bp.

Under our Int-tag assay conditions, the number of concatemerized Int-tags per clone ranged from 3 to 10, with an average of 5 Int-tags/clone. We sequenced a total of 515 Int-tags: 385 were authentic integration site sequences, 27 did not yield a high-quality match to the human genome, and 103 contained viral sequences downstream of the left LTR, which presumably derived from LTR circles. Of the 385 that yielded authentic integration site sequences, 194 (50.4%) contained 19 bp and 191 (49.6%) contained 20 bp of host DNA sequence.

To test the ability of mapping chromosomal locations by use of short DNA sequences, we carried out a simulation experiment with 10,000 randomly selected 19- and 20-bp cellular sequences. The positions of the 22 autosomal chromosomes plus the X sex chromosome were represented by adding their lengths together linearly, and uniformly random positions within the human genome were selected by choosing a random number between 0 and 3,070,128,058 (genome size). We found that 68.4% of these sequences were mapped to unique locations in the human genome (Rdm-tag) (Fig. 5). To further determine the ability of the short sequences to identify chromosomal locations, we generated conventional tags (Cvt-tags) based on the 323 NL-Mme integration sites obtained earlier using the conventional assay. The integration site sequences were divided as randomly and proportionally as the Int-tags (50.4%:49.6%), and the first 19 or 20 nucleotides immediately adjacent to the viral DNA of each integration site sequence were used to map the chromosomal location by use of the BLASTN program. Similarly to the simulation with the computer-generated Rdm-tags, 69.3% of the Cvt-tags were mapped to unique locations (Fig. 5).

The chromosomal locations of the cellular sequences from the Int-tag assay were then mapped using the BLASTN program. We found that, similarly to the Cvt-tags and Rdm-tags, 269 (69.9%) of the 385 authentic integration site sequences were mapped to unique chromosomal locations (Int-tag) (Fig. 5). The *P* value for the comparison of unique location rates for Rdm-tags, Cvt-tags, and Int-tags is 0.665.

For sequences that mapped to two or more chromosomal locations, a majority belonged to one of the several repeat sequence families (32). In all three independently derived tag sequences, although we could not determine the chromosomal location of ~30% of the tag sequences, repeat families (e.g., LINE and *Alu*) associated with 57% of these "multiple hit" sequences could be identified (Fig. 5). Overall, ~70% of 19- and 20-bp tag sequences can be mapped to a unique location in the human genome, while the chromosomal features associated with ~85% of these tag sequences can be identified (Fig. 5).

Validity of the high-throughput Int-tag assay. To determine the validity of the Int-tag assay, we compared the chromosomal features associated with the integration sites as determined by the Int-tag assay with those as determined by the conventional assay (Fig. 6). For additional comparison and to account for potential effects from using short sequences, we also analyzed Rdm-tags and Cvt-tags, which are derived from random cellu-

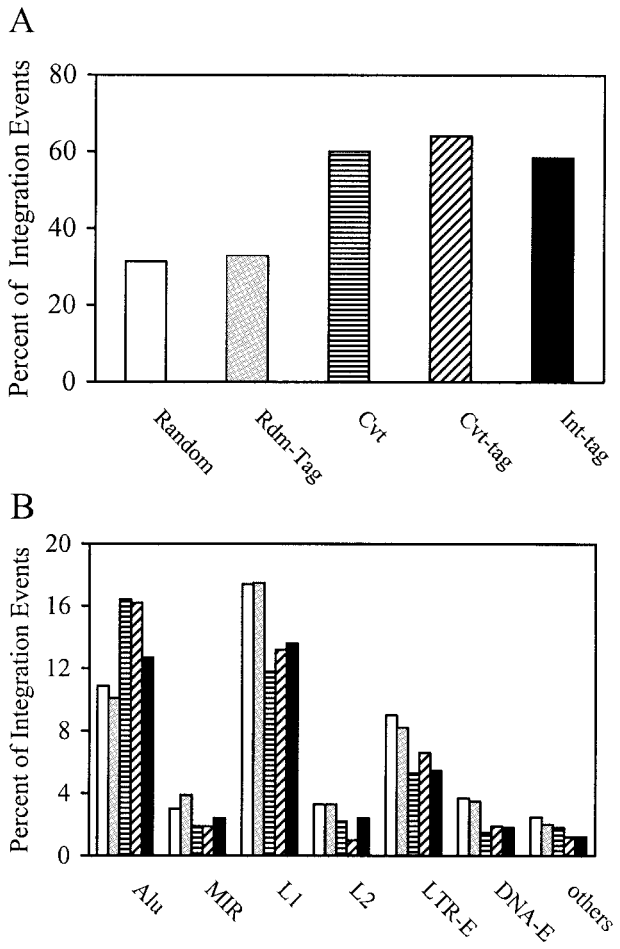

FIG. 6. Analysis of chromosomal features associated with integration events by use of the high-throughput Int-tag assay and the conventional assay. Integration site sequences were determined using the conventional assay (striped bars) or the Int-tag assay (black bars) or generated in silico (open bars). Rdm-tag (gray bars) and Cvt-tag (hatched bars) sequences were derived from integration site sequences generated in silico and by the conventional assay, respectively, as described above. The locations of the integration sites were mapped, and chromosomal features in the vicinity of the integration sites were identified. (A) Transcription units. Integration site sequences that mapped to a unique chromosomal location were analyzed and scored as a part of a transcription unit only if the transcription unit was a member of the RefSeq genes. (B) Identifiable repeat sequences. Integration site sequences that mapped to multiple locations were analyzed for identifiable repeat elements, including *Alu* and mammalian interspersed repeat (MIR) of the SINE, L1 and L2 of the LINE, LTR-E, and DNA-E.

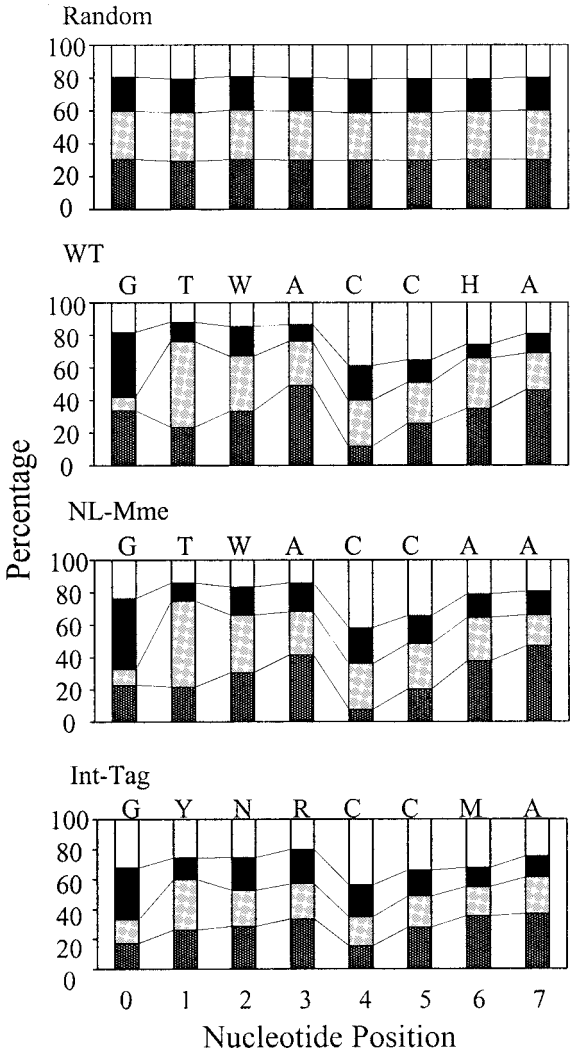

FIG. 7. Base preference in genomic sequence immediately adjacent to integration sites. The numbers on the *x* axis represent nucleotide positions of human DNA adjacent to the proviral DNA, where the point of joining between the HIV and human DNA lies to the left of position 0. The height of the bar represents the percent frequency of each base. A, T, G, and C are denoted by dark gray, light gray, black, and open bars, respectively. The preferred sequence is listed on the top of each panel. H denotes A, C, or T but not G; M denotes A or C; N denotes A, C, G, or T; R denotes A or G; W denotes A or T; and Y denotes C or T.

lar sequences and integration site sequences by use of the conventional assay, respectively. Sequences that mapped to a unique location in the human genome (~70% of total [Fig. 5]) were used for analyzing transcription units (Fig. 6A), and tag sequences that mapped to identifiable chromosomal features (~85% of total [Fig. 5]) were used for analyzing repeat elements (Fig. 6B). In the random control, the distribution of computer-generated integration sites in transcription units (31.4%) and various repeat elements paralleled the relative levels of abundance of the elements in the human genome (25). The distribution of the Rdm-tag sequences was similar to that of the uniformly random integration sites ($P = 0.125$), indicating that the use of short tag sequences did not significantly alter the analysis (Fig. 6). In contrast to the random control, the distribution of NL-Mme integration sites determined by the conventional assay was significantly favored in transcription units (60.1%, $P < 0.0001$) and *Alu* elements (16.4%, $P = 0.0035$) and disfavored in LTR-E (5.3%, $P = 0.0193$) and the L1 member of the LINE class (11.8%, $P = 0.0093$) (Fig. 6). Such a pattern of integration distribution was similar to that published previously using similar methods of analysis (37, 42). The distribution of Cvt-tag sequences showed a bias similar to that of their full-length counterparts (*P* value of 0.527 for the comparison of integration patterns for Cvt and

Cvt-tags), further confirming that analyzing chromosomal features associated with integration sites was not affected by using short tag sequences. For integration sites determined by the Int-tag assay, we also found a significant preference for transcription units (58.4%, $P < 0.0001$) and *Alu* elements (12.7%), while LTR-E (5.4%, $P = 0.0333$) and the L1 repetitive elements (13.6%) were disfavored (Fig. 6). Although the Int-tag assay also detected similar preferences for *Alu* elements (12.7%) and for L1 repetitive elements (13.6%), the results were not statistically different (*P* values are 0.3605 and 0.0808, respectively), probably because of the relatively small sample size. Overall, we did not detect any significant differences in the patterns of integration distribution of NL-Mme between the conventional and Int-tag assays ($P = 0.892$).

As another indicator for evaluating the validity of the Int-tag assay, we analyzed the frequency of each base at positions near the vicinity of the integration site. Although integration of HIV-1 is not associated with specific DNA sequences, a weak consensus sequence of the 5-bp direct repeat at the integration site has been reported previously (8, 40). Examination of the base frequencies directly surrounding a large number of cloned HIV-1 integration sites from infected cells also revealed a preferred integration sequence, 5′-[0]GTWACCHA[7]-3′ (using standard International Union of Biochemistry mixed base codes) (18, 27). The exact sequence was also preferred when the integration site sequences of the WT virus were analyzed (Fig. 7). The integration site sequences of NL-Mme, as determined by the conventional assay, showed similar base preferences (5′-[0]GTWACCAA[7]-3′) in the genomic sequence immediately adjacent to the integration site (Fig. 7). The integration site sequences of NL-Mme determined by the Int-tag assay also showed base preferences, and the preferred sequence (5′-[0]GYNRCCMA[7]-3′) was similar to that determined by the conventional assay (Fig. 7) and to those reported previously (8, 18, 27, 40).

Taken together, the Int-tag assay is a valid and efficient method in determining cellular sequences at the integration site, mapping the chromosomal location, and identifying a particular chromosomal feature associated with the integration event.

DISCUSSION

The new high-throughput assay involves the use of a modified HIV-1, NL-Mme, that contains a type IIS MmeI restriction site at the U5 end of the LTR. This class of enzymes has been used successfully to produce sequence tags for expression profiling (16). When the MmeI-containing viral DNA is integrated into cellular DNA, the new site in the right LTR is positioned such that the MmeI cleavage site is in the cellular DNA. The advantage of this experimental design is that this new cleavage site is always at a fixed distance (19 or 20 bp) from the LTR, regardless of the site of integration. Therefore, digestion with MmeI of cellular DNA from cells infected with NL-Mme generates right-LTR-containing fragments of discrete rather than heterogeneous size. These DNA fragments are converted to Int-tags after ligation-mediated PCR and subsequently concatemerized as a serial tag sequence with the viral and linker sequences serving as punctuation marks (Fig. 1B). Because the size of each Int-tag is relatively small, the concatemerization of these tags allows an efficient and high-throughput analysis of multiple integration sites in a single DNA sequencing reaction. This approach is in contrast to current methods where each provirus-containing clone contains only one integration site and involves sequencing determination of, on average, hundreds of bases (17, 26, 31, 37, 42). The new method also does not employ a selection step to enrich provirus-containing clones or require the presence of certain restriction sites or the repetitive *Alu* element near the provirus (8, 10, 15, 35, 36, 39, 40), which might inadvertently disturb or skew the initial distribution and preference of integration sites.

A full library of 19-bp sequence tags has a complexity of 2.75×10^{11}, which should be sufficient to map any 19-bp tag to a unique address in the human genome of 3.07×10^9 bp. However, the human genome is AT rich (~60%), and at least 50% of the genome consists of repeat sequences (25). Our simulation exercise using a mixture of 10,000 computer-generated 19- and 20-bp sequence tags showed that, even with the uneven distribution of bases and high content of repetitive elements, about 70% of the short tags can be mapped to unique locations in the human genome. By converting the integration site sequences in HIV-infected cells obtained by the conventional assay into 19- and 20-bp tags, we confirmed that 70% of these tag sequences were mapped to the same unique locations. Similarly, 70% of tag sequences derived from the Int-tag assay had a unique chromosomal address.

As expected, a majority of the tag sequences that mapped to multiple locations is associated with repetitive elements. This is consistent with an earlier genome-wide study of 524 HIV-1 integration sites showing that about 30% of proviruses are located near repeat sequences (37). Repetitive elements are grouped into five major classes: (i) transposon-derived (interspersed) repeats, (ii) processed pseudogenes, (iii) simple sequence repeats, (iv) segmental duplications, and (v) pericentromeric and subtelomeric tandem repeats. Over 90% of human repeat sequences are related to or derived from transposable elements, such as LINEs, SINEs, LTR-E, and DNA repeat elements (DNA-E) (32). Although about 30% of the tag sequences, generated either in silico or from the conventional and Int-tag assays, did not provide a unique address, it is important to note that we could classify about half of these tag sequences among the transposon-derived repeats and simple sequence repeats. Therefore, our analyses showed that about 70% of 19- and 20-bp tag sequences can be mapped to unique locations, while 85% can be identified by their chromosomal features.

Genome-wide analysis of integration sites indicates that HIV-1 favors transcription units and *Alu* elements, which are abundant in gene-rich chromosomal domains, and disfavors LTR-E, which are depleted in gene-rich regions of the genome (30, 37, 42). Similar integration site preferences were observed with NL-Mme, indicating that introduction of the MmeI restriction site in the U5 region of the HIV-1 LTR had no measurable effect on the resulting virus. An identical integration preference was also observed when the genome-wide distribution of integration sites of NL-Mme in human cells was analyzed using the Int-tag assay. Furthermore, the preferred integration site sequence as determined by the Int-tags resembles closely those reported previously (8, 18, 40). Therefore,

the Int-tag assay is a valid approach in determining integration site sequences, mapping integration site locations, and identifying chromosomal features associated with the integration event.

Integration of retroviral DNA occurs at many sites within the host cell genome, but the process is not uniformly distributed. The site of integration has significant implications for both the virus and the host cell. Therefore, it is important to gain a better understanding of the distribution and preference of integration sites and factors that affect the site selection process. Although reliable assays have already been established for sequencing and mapping integration sites, studies on integration site choice and its determining factors involve the collection and analysis of numerous libraries, with each consisting of hundreds or thousands of independent integration sites. The availability of the described high-throughput assay will make the process less labor-intensive, less time-consuming, and more cost-effective. In addition to HIV-1, the described methodology can be adapted easily to integration site studies involving other retroviruses and transposons (7, 34).

ACKNOWLEDGMENTS

We thank Joseph L. DeRisi for helpful discussions, Thomas A. Wilkinson for comments and critical reading of the manuscript, and the Core Virology Laboratory at the UCLA AIDS Institute for carrying out the enzyme-linked immunosorbent assay for p24.

This work was supported by National Institutes of Health grant CA68859 and a seed grant from the UCLA AIDS Institute (NIH grant AI28697) to S.A.C.

REFERENCES

1. **Ausubel, F. A., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl.** 1999. Current protocols in molecular biology. Wiley, New York, N.Y.
2. **Balakrishnan, M., and C. B. Jonsson.** 1997. Functional identification of nucleotides conferring specificity to retroviral integrase reactions. J. Virol. **71:**1025–1035.
3. **Barr, S. D., J. Leipzig, P. Shinn, J. R. Ecker, and F. D. Bushman.** 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. J. Virol. **79:**12035–12044.
4. **Brown, P. O.** 1997. Integration, p. 161–204. *In* J. M. Coffin, S. H. Hughes, and H. E. Varmus (ed.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
5. **Bushman, F., M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, and C. Hoffmann.** 2005. Genome-wide analysis of retroviral DNA integration. Nat. Rev. Microbiol. **3:**848–858.
6. **Bushman, F. D.** 2002. Integration site selection by lentiviruses: biology and possible control. Curr. Top. Microbiol. Immunol. **261:**165–177.
7. **Bushman, F. D.** 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell **115:**135–138.
8. **Carteau, S., C. Hoffmann, and F. Bushman.** 1998. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. J. Virol. **72:**4005–4014.
9. **Chow, S. A., and P. O. Brown.** 1994. Substrate features important for recognition and catalysis by human immunodeficiency virus type 1 integrase identified by using novel DNA substrates. J. Virol. **68:**3896–3907.
10. **Chun, T.-W., L. Carruth, D. Finzi, X. Shen, J. A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T. C. Quinn, Y.-H. Kuo, R. Brookmeyer, M. A. Zeigler, P. Barditch-Crovo, and R. F. Siliciano.** 1997. Quantitation of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature **387:**183–188.
11. **Ciuffi, A., M. Llano, E. Poeschla, C. Hoffmann, J. Leipzig, P. Shinn, J. R. Ecker, and F. Bushman.** 2005. A role for LEDGF/p75 in targeting HIV DNA integration. Nat. Med. **11:**1287–1289.
12. **Crise, B., Y. Li, C. Yuan, D. R. Morcock, D. Whitby, D. J. Munroe, L. O. Arthur, and X. Wu.** 2005. Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1. J. Virol. **79:**12199–12204.
13. **Cullen, B. R.** 1995. Regulation of HIV gene expression. AIDS **9**(Suppl. A)**:**S19–S32.
14. **Dao-Thi, M. H., L. Van Melderen, E. De Genst, H. Afif, L. Buts, L. Wyns, and R. Loris.** 2005. Molecular basis of gyrase poisoning by the addiction toxin CcdB. J. Mol. Biol. **348:**1091–1102.
15. **Gaur, M., and A. D. Leavitt.** 1998. Mutations in the human immunodeficiency virus type 1 integrase D,D(35)E motif do not eliminate provirus formation. J. Virol. **72:**4678–4685.
16. **Gnatenko, D. V., J. J. Dunn, S. R. McCorkle, D. Weissmann, P. L. Perrotta, and W. F. Bahou.** 2003. Transcript profiling of human platelets using microarray and serial analysis of gene expression. Blood **101:**2285–2293.
17. **Han, Y., K. Lassen, D. Monie, A. R. Sedaghat, S. Shimoji, X. Liu, T. C. Pierson, J. B. Margolick, R. F. Siliciano, and J. D. Siliciano.** 2004. Resting $CD4^+$ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. J. Virol. **78:**6122–6133.
18. **Holman, A. G., and J. M. Coffin.** 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc. Natl. Acad. Sci. USA **102:**6103–6107.
19. **Holmes-Son, M. L., R. S. Appa, and S. A. Chow.** 2001. Molecular genetics and target site specificity of retroviral integration. Adv. Genet. **43:**33–69.
20. **Jordan, A., D. Bisgrove, and E. Verdin.** 2003. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. EMBO J. **22:**1868–1877.
21. **Jordan, A., P. Defechereux, and E. Verdin.** 2001. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. EMBO J. **20:**1726–1738.
22. **Karn, J.** 1999. Tackling Tat. J. Mol. Biol. **293:**235–254.
23. **Labrador, M., and V. G. Corces.** 2002. Setting the boundaries of chromatin domains and nuclear organization. Cell **111:**151–154.
24. **LaFemina, R. L., P. L. Callahan, and M. G. Cordingley.** 1991. Substrate specificity of recombinant human immunodeficiency virus integrase protein. J. Virol. **65:**5624–5630.
25. **Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al.** 2001. Initial sequencing and analysis of the human genome. Nature **409:**860–921.
26. **Laufs, S., B. Gentner, K. Z. Nagy, A. Jauch, A. Benner, S. Naundorf, K. Kuehlcke, B. Schiedlmeier, A. D. Ho, W. J. Zeller, and S. Fruehauf.** 2003. Retroviral vector integration occurs in preferred genomic targets of human bone marrow-repopulating cells. Blood **101:**2191–2198.
27. **Lewinski, M. K., D. Bisgrove, P. Shinn, H. Chen, C. Hoffmann, S. Hannenhalli, E. Verdin, C. C. Berry, J. R. Ecker, and F. D. Bushman.** 2005. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. J. Virol. **79:**6610–6619.
28. **Masuda, T., V. Planelles, P. Krogstad, and I. S. Y. Chen.** 1995. Genetic analysis of human immunodeficiency virus type 1 integrase and the U3 *att* site: unusual phenotype of mutants in the zinc finger-like domain. J. Virol. **69:**6687–6696.
29. **Mitchell, R., C. Y. Chiang, C. Berry, and F. Bushman.** 2003. Global analysis of cellular transcription following infection with an HIV-based vector. Mol. Ther. **8:**674–687.
30. **Mitchell, R. S., B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker, and F. D. Bushman.** 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. **2:**1127–1137.
31. **Narezkina, A., K. D. Taganov, S. Litwin, R. Stoyanova, J. Hayashi, C. Seeger, A. M. Skalka, and R. A. Katz.** 2004. Genome-wide analyses of avian sarcoma virus integration sites. J. Virol. **78:**11656–11663.
32. **Prak, E. T., and H. H. Kazazian, Jr.** 2000. Mobile elements and the human genome. Nat. Rev. Genet. **1:**134–144.
33. **Reicin, A. S., G. Kalpana, S. Paik, S. Marmon, and S. Goff.** 1995. Sequences in the human immunodeficiency virus type 1 U3 region required for in vivo and in vitro integration. J. Virol. **69:**5904–5907.
34. **Roe, T., S. A. Chow, and P. O. Brown.** 1997. 3′-end processing and kinetics of 5′-end joining during retroviral integration in vitro. J. Virol. **71:**1334–1340.
35. **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.
36. **Scherdin, U., K. Rhodes, and M. Breindl.** 1990. Transcriptionally active genome regions are preferred targets for retrovirus integration. J. Virol. **64:**907–912.

37. **Schroder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman.** 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell **110:**521–529.
38. **Sherman, P. A., M. L. Dickson, and J. A. Fyfe.** 1992. Human immunodeficiency virus type 1 integration protein: DNA sequence requirements for cleaving and joining reactions. J. Virol. **66:**3593–3601.
39. **Shih, C.-C., J. P. Stoye, and J. M. Coffin.** 1988. Highly preferred targets for retrovirus integration. Cell **53:**531–537.
40. **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.
41. **Strebel, K.** 2003. Virus-host interactions: role of HIV proteins Vif, Tat, and Rev. AIDS **17**(Suppl. 4)**:**S25–S34.
42. **Wu, X., Y. Li, B. Crise, and S. M. Burgess.** 2003. Transcription start regions in the human genome are favored targets for MLV integration. Science **300:**1749–1751.