JOURNAL OF VIROLOGY, Feb. 2004, p. 1301–1313 Vol. 78, No. 3
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.3.1301–1313.2004

Fusion Proteins Consisting of Human Immunodeficiency Virus Type 1 Integrase and the Designed Polydactyl Zinc Finger Protein E2C Direct Integration of Viral DNA into Specific Sites

Wenjie Tan,[1] Kai Zhu,[1] David J. Segal,[2]† Carlos F. Barbas III,[2] and Samson A. Chow[1]*

Department of Molecular and Medical Pharmacology, Molecular Biology Institute, and UCLA AIDS Institute, UCLA School of Medicine, Los Angeles, California 90095,[1] and The Skaggs Institute for Chemical Biology and the Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037[2]

Received 19 August 2003/Accepted 14 October 2003

In order to establish a productive infection, a retrovirus must integrate the cDNA of its RNA genome into the host cell chromosome. While this critical process makes retroviruses an attractive vector for gene delivery, the nonspecific nature of integration presents inherent hazards and variations in gene expression. One approach to alleviating the problem involves fusing retroviral integrase to a sequence-specific DNA-binding protein that targets a defined chromosomal site. We prepared proteins consisting of wild-type or truncated human immunodeficiency virus type 1 (HIV-1) integrase fused to the synthetic polydactyl zinc finger protein E2C. The purified fusion proteins bound specifically to the 18-bp E2C recognition sequence as analyzed by DNase I footprinting. The fusion proteins were catalytically active and biased integration of retroviral DNA near the E2C-binding site in vitro. The distribution was asymmetric, and the major integration hot spots were localized within a 20-bp region upstream of the C-rich strand of the E2C recognition sequence. Integration bias was not observed with target plasmids bearing a mutated E2C-binding site or when HIV-1 integrase and E2C were added to the reaction as separate proteins. The results demonstrate that the integrase-E2C fusion proteins offer an efficient approach and a versatile framework for directing the integration of retroviral DNA into a predetermined DNA site.

Integration of the cDNA copy of the retroviral genome into the chromosome of a susceptible host cell is obligatory for retroviruses to establish a productive infection (for a review, see reference 6). Integration is catalyzed by the virus-encoded protein integrase and proceeds through a highly ordered multistep process resulting in the formation of a provirus. The provirus becomes part of the cellular genome for the lifetime of the infected cell and allows the retrovirus to employ the cellular transcriptional and translational machinery for synthesizing viral progeny.

A salient feature of retroviral integration is that the site of insertion can occur throughout the chromosomes of the target cell. Analyses of integration sites in cells infected with human immunodeficiency virus type 1 (HIV-1) (9, 48), avian sarcoma virus (25), avian leukosis virus (56), and Rous sarcoma virus (52) revealed that most regions of the cellular DNA are accessible. However, the same region or exact nucleotide sequence in the host cell genome can be utilized at a frequency several hundred-fold greater than chance, lending credence to the idea that there are hot and cold spots for integration (14, 48, 56). Therefore, integration of retroviral DNA into target DNA is nonspecific, but it is not a random process.

The ability of retroviruses to permanently insert their genome into the chromosome of an infected cell is a property that can be exploited for gene therapy (for reviews, see references 30 and 42). However, because the integration reaction is an inherently mutagenic process, the nonspecific nature of integration can be a potential pitfall for introducing a transgene with retroviral vectors (for reviews, see references 53 and 54). Depending on the site of integration, insertional mutagenesis may disrupt normal cell functions by inactivating an essential host gene or inappropriately causing overexpression of an undesirable gene, such as a proto-oncogene. Development of leukemia associated with the use of retroviral vectors in gene therapy trials has been reported (22, 36), but the level of risk of cancer and other side effects caused by insertional mutagenesis has not been assessed carefully.

One strategy to control the site specificity of retroviral integration is through the use of a fusion protein consisting of a retroviral integrase and a sequence-specific DNA-binding protein, such as phage λ repressor (7), *Escherichia coli* LexA repressor (20, 29), and murine transcription factor Zif268 (8). In vitro, the sequence-specific DNA-binding proteins direct integration by recognizing and binding to their target sites on the DNA, causing integration to be mediated into the adjacent regions. A major limitation of the strategy is that the DNA-binding sequences of the previously tested fusion proteins are defined and fixed and may not necessarily be localized to a desired chromosomal site. In addition these DNA-binding proteins can recognize multiple DNA variants of their consensus binding sequence, or the number of nucleotides required for specific protein-DNA interaction is insufficient for specifying a unique site within a mammalian genome (26, 35, 41). Using the HIV-1 integrase-LexA fusion protein as an example, although LexA protein binds to a 16-bp sequence with approximate

* Corresponding author. Mailing address: UCLA School of Medicine, 23-133 CHS, 10833 Le Conte Ave., Los Angeles, CA 90095. Phone: (310) 825-9600. Fax: (310) 825-6267. E-mail: schow@mednet.ucla.edu.

† Present address: Department of Pharmacology and Toxicology, College of Pharmacy, University of Arizona, Tucson, AZ 85721.

TABLE 1. DNA sequences of PCR primers and oligonucleotides used in construction of fusion proteins and DNA substrates for integration assays

Primer	Sequence[a]	Restriction enzyme
E2C(+)	5′-ATATAGGTACCTTGGCCCAGGCGGCCCTCGAG-3′	*Kpn*I
E2C(−)	5′-ATTGGATCCTTACTGGCCGGCCTGGCCACT-3′	*Bam*HI
E2CF1	5′-ATATTCATATGTGGGCCCAGGCGGCCCTCGAG-3′	*Nde*I
E2CR1	5′AAACGGGTACCGGCCGGCCTGGCCACTAGTT-3′	*Kpn*I
INF	5′-GCCGGCCGGTACCCGTTTTTAGATGGAATAG-3′	*Kpn*I
INFΔ10	5′-GCCGGCCGGTACCCGGAACATGAGAAATATCACAG-3′	*Kpn*I
INR	5′-ATTGGATCCTCAATCCTCATCCTGTCTACT-3′	*Bam*HI
GST(+)	5′-GCGTGGTACCATATGTCCCCTATACTAGG-3′	*Kpn*I
GST(−)	5′-TCAGCTGCAGTCAATCCGATTTTGGAGGATGG-3′	*Pst*I
Te2c(+)[b]	5′-AGCTTGGTGCT*CACTGCGGCTCCGGCCCC*ATG-3′	
Te2c(−)[b]	5′-GATCCAT*GGGGCCGGAGCCGCAGTG*AGCACCA-3′	
mTe2c(+)[b]	5′-AGCTTGGTGCT*CACTGCGGCTC**TCGAACT***ATG-3′	
mTe2c(−)[b]	5′-GATCCAT*AG**TTCGA**GAGCCGCAGTG*AGCACCA-3′	
B2-1	5′-ATGTGGAAAATCTCTAGCA-3′	
C220	5′-ATGTGGAAAATCTCTAGCAGT-3′	
V2	5′-ACTGCTAGAGATTTTCCACAT-3′	
PR-G	5′-CACAGGAAACAGCTATGACCATG-3′	
PR-C	5′-CACGACGTTGTAAAACGACGGCC-3′	

[a] Restriction sites are underlined, with the identity of the restriction enzyme listed to the right.

[b] Italic letters denote the E2C-binding sequence, and nucleotide substitutions resulting in the mutant E2C-binding site are identified in bold type.

twofold rotational symmetry, only three nucleotides at each end of the palindrome are highly conserved among the binding sties (35). In the human genome of 3 × 10^9 bp, we estimated that there are thousands of potential LexA-binding sites. The relatively low binding specificity, coupled with the difficulty of incorporating the fusion protein into infectious virions (8, 29), have made it difficult to assess whether these fusion proteins are able to direct integration in vivo.

One class of DNA-binding proteins that offers several advantages in conferring site specificity to retroviral integrases are the synthetic proteins derived from the Cys_2-His_2 zinc finger proteins (for reviews, see references 3 and 49). Structural studies of the Cys_2-His_2 zinc finger domain showed that it has a simple ββα fold of ~30 amino acids in length and is stabilized by hydrophobic interactions and zinc chelation (34, 39). Analysis of the three-zinc-finger protein Zif268–DNA complex revealed that the α-helix of each zinc finger fits directly into the major groove and the amino acid side chains make specific contacts with a 3-bp DNA subsite. Most of the base contacts involve the G-rich strand of the binding site (18, 41). Studies directed at modifying the sequence specificity of the zinc finger DNA-binding domains have shown that they can be selected to specifically bind a wide array of DNA sequences. In addition, many selected zinc finger domains exhibit sufficient modularity in their recognition of DNA triplets that they can be combined with other such domains to create polydactyl proteins that recognize extended sequences of DNA (5, 11, 17, 31, 37, 38, 46, 47, 50). One example of such a polydactyl protein is E2C, which contains six zinc finger domains. E2C was constructed by grafting the amino acid residues of each zinc finger involved in specific DNA recognition into the framework of the designed consensus protein Sp1C, a derivative of Sp1 (15). E2C binds with high affinity and recognizes a contiguous 18-bp sequence, which is unique in the human genome and located within the 5′ untranslated region of the *erbB-2* gene on chromosome 17 (4, 5). Artificial transcription factors based on modified zinc finger domains have been used to target specific DNA sequences and selectively activate or repress expression of reporter genes (4, 5, 10, 12, 16, 28, 32, 33, 37, 47).

Herein, we constructed and purified various fusion proteins consisting of HIV-1 integrase and the polydactyl zinc finger protein E2C. The fusion proteins retained their integration activity and ability to bind specifically to the E2C-binding site. Analysis of the distribution and frequency of integration events revealed that integration of viral DNA mediated by the integrase-E2C fusion proteins was biased near the E2C-binding site.

MATERIALS AND METHODS

DNA constructs. For the fusion protein IN1-288/E2C, which has the polydactyl zinc finger protein E2C fused to the C terminus of the full-length HIV-1 integrase, the plasmid containing the E2C gene, pMal-c2-E2C (5), was amplified by PCR using E2C(+) and E2C(−) as the forward and reverse primers, respectively. Oligonucleotides used in PCR were from Operon Technologies, Inc. (Alameda, Calif.), and the sequences are shown in Table 1. The amplification generated a fragment encoding all 184 amino acid residues of E2C. The DNA fragment was digested with *Kpn*I and *Bam*HI and purified with the QIAEX gel extraction kit (Qiagen, Chatsworth, Calif.). The purified *Kpn*I-*Bam*HI fragment was then ligated in frame with pIN1-288/LA previously digested with *Kpn*I and *Bam*HI to form pIN1-288/E2C. pIN1-288/LA is derived from the pT7-7 expression vector and contains the fusion gene encoding HIV-1 integrase and the *E. coli* LexA protein (20). The *lexA* gene in the construct is flanked by the restriction enzymes *Kpn*I and *Bam*HI. The described procedure replaced the *lexA* gene with *e2c*. Similarly, the plasmid constructs for the fusion protein consisting of N-terminal-truncated HIV-1 integrase and E2C, pIN50-288/E2C, and the fusion construct containing the C-terminal-truncated HIV-1 integrase and E2C, pIN1-234/E2C, were prepared by ligating the *Kpn*I-*Bam*HI PCR fragment to *Kpn*I-*Bam*HI-digested pIN50-288/LA and pIN1-234/LA, respectively (20).

Two additional fusion constructs, E2C/IN1-288 and E2C/IN11-288, were prepared in which the polydactyl zinc finger protein E2C was placed at the N terminus of HIV-1 integrase. E2C/IN1-288 contains a full-length HIV-1 integrase, whereas the integrase in IN/11-288 has a 10-amino-acid deletion at the N terminus. The fusion genes were obtained by an overlapping PCR method that generated a 5′ and 3′ fragment with a shared region of homology (2). In the first set of PCRs for E2C/IN1-288, the 5′ fragment was obtained using pMal-c2-E2C as the template and E2CF1 and E2CR1 as the forward and reverse primers, respectively. The 3′ fragment was amplified using pT7-7/H-IN (20) as the template and INF and INR as the forward and reverse primers, respectively. An identical procedure was used for E2C/IN11-288, except the 3′ fragment was obtained using INFΔ10 and INR as the forward and reverse primers, respectively. The two PCR products containing a common overlapping region were annealed together, extended, and amplified in the presence of 0.2 mM deoxyribonucleoside triphosphates (United States Biochemicals), 2.5 U of *Pfu* polymerase (*Pfu* Turbo; Stratagene), and E2CF1 and INR as the forward and reverse primers, respectively. The extension and amplification were carried out in a thermocycler (MJ Research, Inc.) programmed for three cycles, with each cycle consisting of 5 min of denaturation at 95°C, 1 min of annealing at 50°C, and 2 min of extension at 72°C. This was followed by an additional 30 cycles, with each cycle consisting of 40 s of denaturation at 95°C, 40 s of annealing at 58°C, and 2 min of extension at 72°C. The final PCR products from the two separate reactions were digested with *Nde*I and *Bam*HI, gel purified, and then ligated to pT7-7/H-IN previously cut with *Nde*I and *Bam*HI to form pE2C/IN1-288 and pE2C/IN11-288, respectively.

As a control, the E2C protein was also prepared as a fusion with glutathione *S*-transferase (GST), which was inserted at the C terminus of E2C. The GST gene was amplified by PCR using pGEX-2T (Amersham Biosciences, Piscataway, N.J.) as the template and GST(+) and GST(−) as the forward and reverse primers, respectively. The amplified product was digested with *Kpn*I and *Pst*I, followed by ligation to pE2C/IN1-288 previously digested with *Kpn*I and *Pst*I to remove the entire integrase gene.

To prepare a plasmid that contained a single binding site for the E2C protein, a double-stranded oligonucleotide containing the E2C-binding sequence, formed by annealing oligonucleotides Te2c(+) with Te2c(−), was inserted into the *Hin*dIII and *Bam*HI sites of a plasmid derived from pBluescript II KS(+) (Stratagene), resulting in pBS-e2c. To prepare a plasmid that contained a mutant binding site for the E2C protein, two oligonucleotides, mTe2c(+) and mTe2c(−), which contain five nucleotide mutations in the E2C-binding sequence, were annealed and inserted into the identical location as described earlier for pBS-e2c to form pBS-e2cm.

The sequences of all the PCR-amplified DNA fragments were verified by restriction enzyme analysis and the dideoxynucleotide chain termination method. Sequencing reactions were carried out by the UCLA Sequencing Core Facility using an ABI 3700 DNA analyzer (PE Applied Biosystems, Foster City, Calif.).

Expression and purification of the fusion proteins. Wild-type and integrase-E2C fusion proteins were expressed and purified using a protocol similar to that described previously (1). The DNA constructs were transformed into *E. coli* BL21(DE3) cells. The bacterial colony was grown for 16 h at 37°C in 50 ml of Luria broth containing 80 μg of ampicillin (LB-amp)/ml. Forty milliliters of the bacterial culture was used to inoculate 2 liters of prewarmed LB-amp at 32 or 35°C. Expression of the recombinant protein was induced by the addition of 0.4 mM isopropyl-1-thio-β-D-galactopyranoside when the optical density at 600 nm reached 0.8 to 1.0, and the culture was grown for an additional 3 h. The cell pellet from every liter of culture was resuspended in 40 ml of cold lysis buffer (20 mM HEPES [pH 7.5], 10% glycerol, 5 mM 2-mercaptoethanol, 2 μg of leupeptin/ml, 1 mM phenylmethylsulfonyl fluoride, 1 M NaCl, 0.2 mM EDTA, 0.5% Igepal, and 0.2 μg of hen egg lysozyme/ml). The cell suspensions were sonicated three times in a 1-min ice-water bath using a 0.5-inch horn tip (Branson Sonifier 450). The lysate was then centrifuged at 100,000 × *g* for 1 h at 4°C. The supernatant fraction, after dialysis against buffer A (20 mM HEPES [pH 7.5], 10% glycerol, 5 mM 2-mercaptoethanol, 1 M NaCl, 0.1% Igepal), was mixed with Ni^{2+}-nitrilotriacetic acid agarose resin (Qiagen) on ice for 2 h and then washed four times with 10 ml of buffer A containing 50 mM imidazole. The resin was packed into a 15-cm by 0.7-cm (inner diameter) Econo-Column (Bio-Rad), and protein was eluted by applying 15 ml of buffer A with a linear gradient of 50 to 500 mM imidazole at 1 ml/min. The fractions containing the protein were pooled and dialyzed against buffer C {20 mM HEPES (pH 7.5), 20% glycerol, 1 mM dithiothreitol (DTT), 0.5 M NaCl, 0.1 mM EDTA, and 10 mM 3-[(3-cholamidopropyl)-dimethylammonio]-1-propanesulfonic acid (CHAPS)}. The His tag preceding the various fusion proteins was removed by incubating the protein with 80 to 100 NIH U of human thrombin (Sigma) per mg of protein and passing the digested protein through a cation-exchange chromatography column (5 cm by 1 cm [inner diameter]) packed with high-performance SP-Sepharose resin (Amersham Pharmacia). The protein solution was diluted in buffer D (20 mM HEPES [pH 7.0], 10% glycerol, 10 mM DTT, 0.1 mM EDTA, 10 mM CHAPS) to 0.1 mg/ml before loading onto the SP-Sepharose column. A gradient from 0 to 1 M NaCl in buffer D was used to elute the protein from the column. Peak fractions containing the non-His-tagged protein were pooled, concentrated by Centricon-10 columns (Amicon), and dialyzed against buffer GF (20 mM HEPES [pH 7.5], 10% glycerol, 1 mM DTT, 0.5 M NaCl, 0.1 mM EDTA, 10 mM CHAPS). The dialysate was then applied to a HiPrep 26/60 Sephacryl S-200 high-resolution gel filtration column (Amersham Pharmacia) previously equilibrated with buffer GF. The protein was eluted with buffer GF at a flow rate of 0.1 ml/min at 4°C on a BioLogic workstation system (Bio-Rad). Peak fractions containing the full-length protein were pooled and concentrated by a Centricon-10 column or using a stirred cell (model 8050; Amicon) with a YM10 ultrafiltration membrane (Millipore) at a N_2 pressure of 50 lb/in^2. The protein was then dialyzed against storage buffer (20 mM HEPES [pH 7.5], 20% glycerol, 50 μM $ZnCl_2$, 0.3 M NaCl, 10 mM DTT, 10 mM CHAPS) overnight and stored at −80°C. Protein concentrations were determined by the Bradford assay (Bio-Rad) according to the manufacturer's instructions, using bovine serum albumin (BSA) as a standard.

For the E2C-GST fusion protein, the initial steps of protein expression and purification were identical to those described earlier for the integrase-E2C fusion protein. After incubation with human thrombin to remove the His tag at the N terminus, the digested protein was passed through a glutathione-Sepharose affinity chromatography column (Amersham Pharmacia). The column was washed with 10 ml of a buffer containing a final concentration of 10 mM sodium phosphate (pH 7.4), 1 M NaCl, and 0.1% Triton X-100, and the protein was eluted with glutathione elution buffer (100 mM Tris-HCl [pH 7.5], 0.5 M NaCl, 10 mM glutathione, 0.1% Triton X-100). After elution, the protein was dialyzed against storage buffer as described earlier.

Footprinting analysis of DNA binding. To examine the ability of the various E2C-integrase fusion proteins to specifically recognize the E2C-binding site, pBS-e2p, which contains a single E2C-binding sequence, was digested with *Xho*I. The linearized DNA was labeled at the 3′ ends using [α-^{32}P]dCTP and Klenow fragment of *E. coli* DNA polymerase I (New England Biolabs). The labeled DNA was then digested with *Pst*I, and the 382-bp singly end-labeled fragment containing the E2C-binding sequence was isolated from a 1% agarose gel with a QIAEX gel extraction kit (Qiagen). The labeled strand contained the G-rich sequence of the E2C-binding site. To analyze the DNase I digestion pattern of the C-rich sequence of the E2C-binding site, pGL3basic-e2p DNA containing an E2C-binding site (5) was digested with *Nco*I. The linearized DNA was labeled at the 5′ ends using [γ-^{32}P]ATP and T4 polynucleotide kinase. The labeled DNA was digested with *Pst*I, and the 395-bp singly end-labeled fragment containing the E2C-binding sequence was isolated from a 1% agarose gel as described previously. The 3′ or 5′ singly end-labeled DNA fragment (0.3 nM) was incubated with or without protein at room temperature for 30 min in a buffer containing a final concentration of 20 mM HEPES (pH 7.5), 0.05% Igepal, 1.5 mM $CaCl_2$, 2.5 mM $MgCl_2$, 50 mM NaCl, 10 mM DTT, 100 μg of BSA/ml, and 2 μg of poly(dI-dC)/ml. The sample was digested with 2 ng of DNase I/ml for 3 min at room temperature. The digestion was stopped by the addition of 18 mM EDTA, and the sample was deproteinized by phenol-chloroform extraction, ethanol precipitated in the presence of 10 μg of tRNA as a carrier, and resuspended in 5 μl of formamide-10 mM EDTA. After denaturation at 90°C for 3 min, the sample was analyzed by electrophoresis through a 6% denaturing polyacrylamide gel containing 7 M urea in a Tris-borate-EDTA buffer.

In vitro assays for integrase activity. In vitro activities of HIV-1 integrase and the various integrase-E2C fusion proteins were determined using established oligonucleotide-based assays (13). The 3′-end-processing and 3′-end-joining (strand transfer) reactions were carried out at 37°C for 1 h in a 20-μl reaction volume containing 5 nM ^{32}P-labeled substrate, 75 nM purified enzyme, 20 mM HEPES (pH 7.5), 30 mM NaCl, 10 mM $MnCl_2$, 10 mM DTT, and 0.05% Igepal. The oligonucleotides used as DNA substrates were purified by electrophoresis through a 15% denaturing polyacrylamide gel. Oligonucleotides B2-1 and C220 were labeled at the 5′ end with [γ-^{32}P]ATP and T4 polynucleotide kinase (New England Biolabs). The substrate used to assay the 3′-end-processing and 3′-end-joining activities was a double-stranded oligonucleotide containing sequences derived from the U5 end of the HIV-1 long terminus repeat. The substrate was prepared by annealing the labeled C-220 strand with its complementary oligonucleotide V2. To assay only the 3′-end-joining activity, a substrate that resembles the viral U5 end after 3′-end processing was used. The preprocessed substrate was prepared by annealing the labeled B2-1 strand with the V2 strand. The reaction was stopped by adding 18 mM EDTA and heating at 90°C for 3 min before analysis by electrophoresis on a denaturing 15% polyacrylamide gel with 7 M urea in Tris-borate-EDTA buffer. The gel was then dried and placed into a PhosphorImager cassette (Molecular Dynamics), and the reaction products were analyzed with the ImageQuant software (Molecular Dynamics).

PCR-based assay for distribution and frequency of integration events. The PCR-based integration assay is used for analyzing target DNA sites chosen for integration (44, 51). Individual integration events along a target DNA are amplified through PCR to show the distribution and frequency of integration. One microgram of target DNA, pBS-e2c or pBS-e2cm, was preincubated with wild-type HIV-1 integrase or the fusion protein at room temperature for 15 min in a

20-μl final volume of the standard reaction buffer. The donor substrate was the 21-bp oligonucleotide mimicking the preprocessed HIV-1 U5 end and was prepared by annealing B2-1 to V2. The integration reaction was started by adding 15 nM preprocessed U5 DNA and incubating at 37°C for 30 or 60 min. The reaction was terminated with the addition of 80 μl of stop solution (10 mM Tris-HCl [pH 7.5], 5 mM EDTA [pH 8.0], 375 mM sodium acetate, 0.25 mg of tRNA/ml). The DNA was extracted with phenol-chloroform, ethanol precipitated, and resuspended in 50 μl of 10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 8.0). A 3-μl aliquot of the reaction mixture was then added to a buffer containing 10 mM Tris-HCl (pH 8.8), 50 mM KCl, 0.001% gelatin (wt/vol), 1.5 mM $MgCl_2$, 200 μM deoxyribonucleoside triphosphates, 5 pmol each of the forward and reverse primers, and 1 U of *Taq* polymerase (*Taq* 2000; Stratagene) in a final volume of 20 μl. To monitor integration events occurring into the DNA strand containing the G-rich sequence of the E2C-binding site, 0.25 μM oligonucleotide PR-G was used as the reverse primer. The forward primer for the PCR, B2-1, is complementary to the U5 donor substrate and was prepared by mixing 0.05 μM 5′-end-labeled B2-1 and 0.20 μM unlabeled B2-1. To monitor integration events occurring into the DNA strand containing the C-rich sequence of the E2C-binding site, 0.25 μM oligonucleotide PR-C and B2-1 were used as the reverse and forward primers, respectively. Integration events were amplified by 25 or 35 cycles of PCR: 1 min at 94°C, 1 min at 55°C, and 2 min at 72°C. The radiolabeled PCR products were resolved on a 6% denaturing polyacrylamide gel containing 7 M urea in a Tris-borate-EDTA buffer and analyzed with a PhosphorImager (Molecular Dynamics).

RESULTS

Preparation of purified fusion proteins and determination of catalytic activity. The primary structure of various fusion proteins consisting of HIV-1 integrase and the polydactyl zinc finger protein E2C is shown in Fig. 1A. Since there is no a priori information on the configuration of the fusion partners for optimal DNA binding and integration activity, fusion proteins containing full-length HIV-1 integrase at the N terminus (IN/E2C) or the C terminus of E2C (E2C/IN) were prepared. Also, in an attempt to increase integration specificity towards the E2C-binding sequence in the target DNA, we prepared several fusion proteins containing HIV-1 integrase with a deletion in the domain known to bind or interact with target DNA. These included a C-terminal deletion of integrase fused to the N terminus of E2C (IN1-234/E2C) (45, 55, 57) and an N-terminal deletion of integrase fused to the N terminus (IN50-288/E2C) or C terminus (E2C/IN11-288) of E2C (23). A fusion protein consisting of E2C linked at its C terminus to GST (E2C/GST) was prepared as a control for specific binding to the E2C recognition sequence, but it was unable to catalyze integration.

All chimeric integrases were expressed in *E. coli* with a poly-His tag at the N terminus and purified using nickel-chelate affinity chromatography. The His tag was removed by thrombin digestion followed by a second purification step using both cation exchange and gel filtration chromatography. A Coomassie blue-stained sodium dodecyl sulfate (SDS)-polyacrylamide gel showed that the apparent masses of all proteins were consistent with those predicted by the amino acid sequence, ranging from 46 to 52 kDa (Fig. 1B).

The activities of the fusion proteins were first tested for their abilities to catalyze 3′-end processing and 3′-end joining using oligonucleotide-based assays (13). A representation of the in vitro activity is shown in Fig. 2, and the results are summarized in Table 2. In this assay, 3′-end-processing and -joining activities are assayed by the appearance of a product that is shortened by two nucleotides and products that are longer in length than the input DNA, respectively (Fig. 2). The lengths of the 3′-end-joining products are heterogeneous because the site of joining is nonspecific. Fusion of full-length HIV-1 integrase to the N or C terminus of E2C did not change appreciably the 3′-end-processing and -joining activities from those of the wild-

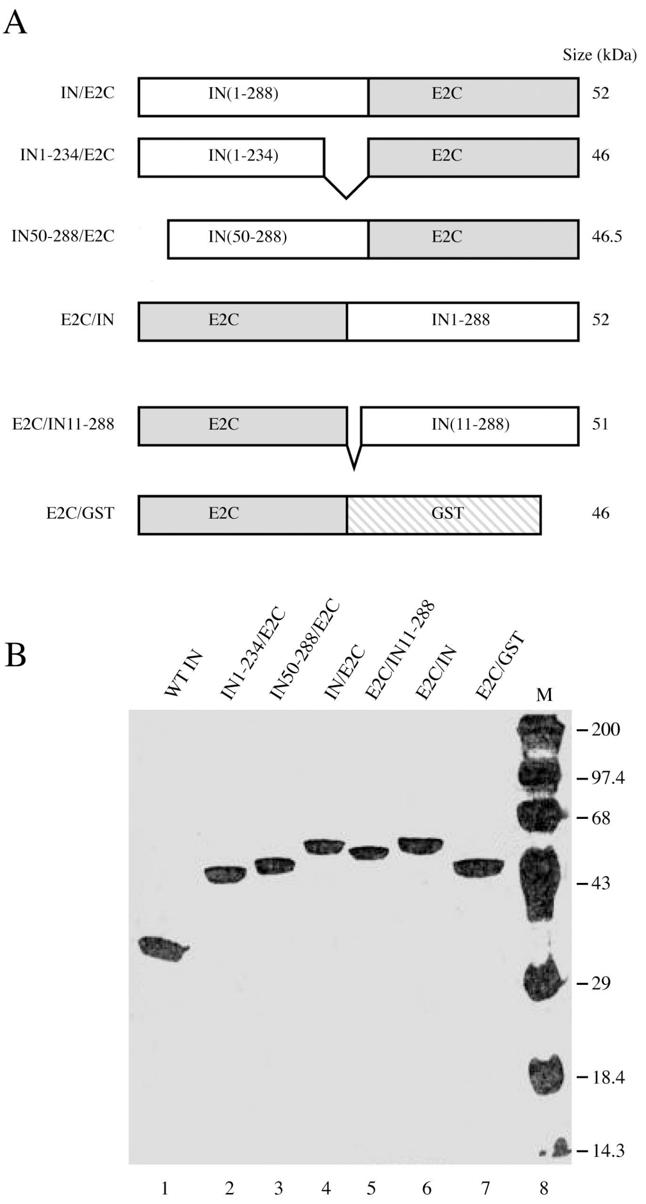

FIG. 1. (A) Primary structures of HIV-1 integrase and E2C fusion proteins. Open and shaded boxes represent the peptides derived from HIV-1 integrase (IN) and polydactyl zinc finger protein E2C, respectively. The stippled box represents peptides from GST. The numbers in parentheses correspond to the amino acid residues included in each fusion protein. Full-length HIV-1 integrase and the E2C protein have 288 and 183 amino acids, respectively. The predicted molecular mass (in kilodaltons) of the various recombinant proteins is indicated on the right. The peptide containing seven consecutive His residues (His tag) used for affinity chromatography was removed from the N terminus by thrombin cleavage during purification. (B) Coomassie blue-stained SDS-polyacrylamide gel of various purified proteins. One microgram of each purified protein as labeled on the top was run on an SDS–12.5% polyacrylamide gel (lanes 1 to 7). Lane 8 contains the molecular weight standards (Gibco BRL) with masses in kilodaltons indicated on the left.

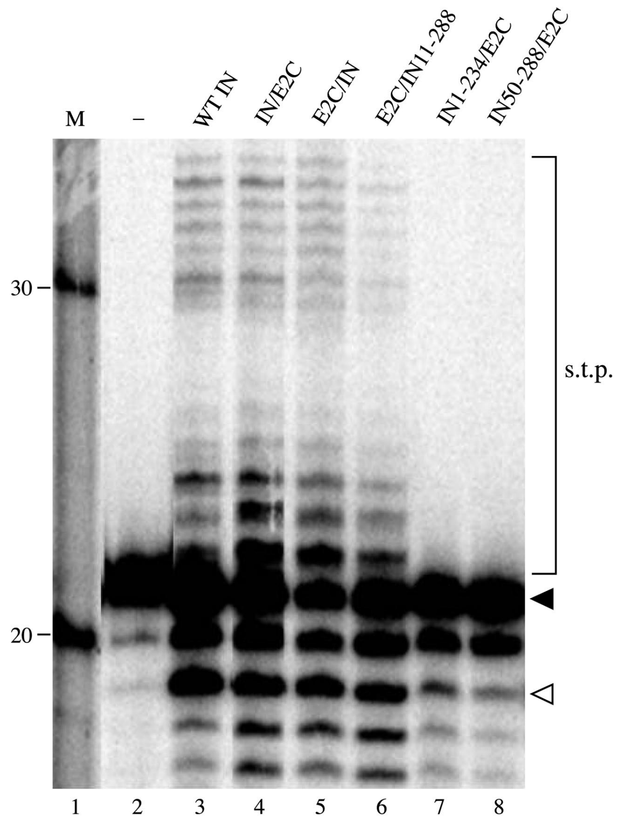

FIG. 2. Catalytic activities of HIV-1 integrase-E2C fusion proteins. The reaction was carried out with 5 nM U5 end oligonucleotide (C220/V2) and a 100 nM concentration of the indicated proteins (lanes 3 to 8). Lane 1 contains the size marker, and the lengths of DNA in nucleotides are indicated on the left. Lane 2 represents a reaction done in the absence of protein. The filled arrowhead denotes the position of the substrate (21-mer), and the open arrowhead indicates the position of the 3′-end-processing product (19-mer). The bands that migrated above the substrate are the products of the 3′-end-joining reaction (strand transfer products [s.t.p.]).

type integrase (Fig. 2, lanes 4 and 5 versus lane 3). A fusion protein consisting of a 10-amino-acid deletion at the N terminus of HIV-1 integrase tethered to the C terminus of E2C retained a wild-type level of 3′-end-processing activity but an approximately 50% decrease in the 3′-end-joining activity (Fig. 2, lane 6). Fusion proteins containing a larger deletion in either the N or C terminus of integrase, IN50-288/E2C or IN1-234/E2C, had a weak 3′-end-processing activity and an undetectable level of 3′-end-joining activity with the oligonucleotide substrates (Fig. 2, lanes 7 and 8).

TABLE 2. Summary of in vitro activities of HIV-1 integrase and E2C fusion proteins

Integrase derivative	Relative activity[a]	
	3′-end processing	3′-end joining
IN/E2C	+++	+++
E2C/IN	+++	+++
E2C/IN11-288	+++	+
IN50-288/E2C	+	+/−[b]
IN1-234/E2C	+	+/−[b]

[a] Expressed by symbols representing the percentage of activity relative to that of wild-type integrase. +++, 50 to 75%; ++, 25 to 50%; +, 25% or less; +/−, less than 5%.

[b] Although little or no 3′-end-joining activity was observed in the oligonucleotide-based assay, strand transfer products were detected in the PCR-based assay (see Fig. 6).

Sequence-specific DNA binding of HIV-1 integrase-E2C fusion proteins. The abilities of the various purified fusion proteins to recognize and bind specifically to an E2C-binding sequence were examined by DNase I footprinting analysis (Fig. 3). When the 3′-end-labeled DNA strand containing the C-rich sequence of the 18-bp E2C-binding site (Fig. 3A) was subjected to DNase I cleavage, addition of wild-type HIV-1 integrase did not display any specific region of protection and the digestion pattern was identical to that obtained in the absence of integrase or fusion proteins (Fig. 3A, compare lanes 3 and 4). The lack of a specific protection from wild-type integrase was expected, because it binds DNA nonspecifically (19, 27, 55, 57). As a positive control, addition of a GST fusion to E2C produced a protected region corresponding in both size and location to the E2C-binding site (Fig. 3A, lane 5). This was consistent with the previous result showing that the C-rich strand of the polydactyl protein-binding site is protected against DNase I digestion (38). Protection of the E2C-binding sequence was also observed with the various E2C-containing fusion proteins (Fig. 3A, lanes 6 to 10), providing direct evidence for sequence-specific DNA binding of these proteins. In addition, the protection afforded by fusion proteins with E2C fused to the C terminus of integrase (Fig. 3A, lanes 8 to 10) was about six nucleotides larger than that afforded by fusion proteins with E2C fused to the N terminus of integrase or GST (Fig. 3A, lanes 5 to 7). The small difference in the size of the protected region may reflect a difference in the mode of DNA binding or protein conformation between the two groups of fusion proteins, resulting in a difference in access to DNase I.

A similar result was obtained when the DNA fragment was singly labeled at the 5′ end of the strand containing the G-rich sequence of the E2C-binding site (Fig. 3B). Comparatively, we noticed that the protein footprints on the G-rich strand (Fig. 3B) were larger in size and better protected than those on the C-rich strand (Fig. 3A). This may be related to the finding that most of the DNA contacts by Zif268 are along the G-rich strand of the binding site (18, 41). Nevertheless, the results showed that tethering E2C with GST or different derivatives of HIV-1 integrase at its N or C terminus did not affect the ability of E2C to recognize and bind its cognate DNA sequence. However, we did not examine whether fusion with integrase alters the binding affinity of E2C to DNA.

Site-directed integration mediated by HIV-1 integrase-E2C fusion protein. A PCR-based assay (Fig. 4A) was used to examine the usage of target sites by HIV-1 integrase and the various fusion proteins (1, 20). Integration reactions were conducted as described in Materials and Methods. The plasmid BS-e2c (Fig. 4B), which contains a single E2C-binding site, was used as the target DNA for integration. The integration products into the target DNA strand containing the C-rich sequence of the E2C-binding site were amplified by PCR and analyzed on a denaturing polyacrylamide gel (Fig. 5A). Each band on the gel corresponds to an integration event at a given phosphodiester bond. The frequency of integration at a particular site and its exact position can be determined by the

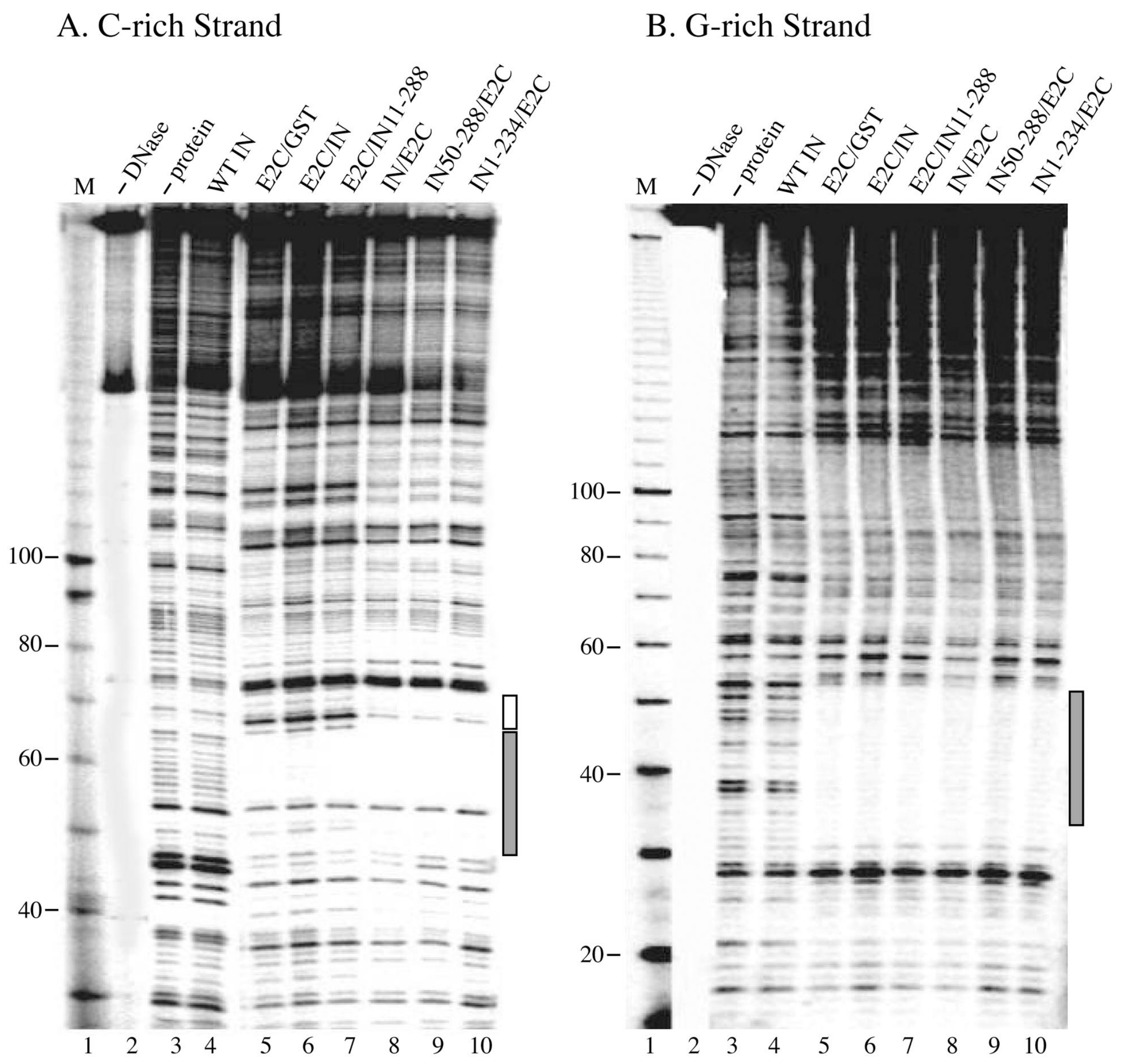

FIG. 3. Footprinting analysis of protein binding to an E2C recognition sequence. (A) Digestion pattern of the C-rich strand of the E2C-binding site. The 395-bp *Pst*I-*Nco*I DNA fragment of pGL3Basic-e2p (5) was singly labeled at the 5′ end of the DNA strand containing the C-rich sequence of the E2C-binding site. (B) Digestion pattern of the G-rich strand of the E2C-binding site. The 382-bp *Xho*I-*Pst*I DNA fragment of pBS-e2p was singly labeled at the 3′ end of the strand containing the G-rich sequence of the E2C-binding site. In both panels, the labeled fragment (0.3 nM) was incubated with a 50 nM concentration of the indicated protein (lanes 4 to 10) for 30 min at room temperature. The samples were then digested with DNase I (2 ng/ml) for 3 min at room temperature, and the digested products were separated on a denaturing polyacrylamide gel. Lane 1 contains size markers, with the DNA lengths in nucleotides indicated on the left. Lane 2 represents the undigested, singly end-labeled DNA fragment. Lane 3 represents a digestion carried out in the absence of protein. The stippled boxes on the right indicate the locations of the E2C-binding site. In panel B, the open box indicates the extended region of protection observed with fusion proteins containing E2C at the C terminus of the integrase (lanes 8 to 10).

intensity of the band and by use of a sequencing ladder, respectively. In reactions using wild-type HIV-1 integrase, the distribution and intensity of PCR-amplified products showed that most positions on the plasmid DNA could be used as target sites for integration, and there was a wide variation in integration frequency among the target sites (Fig. 5A, lanes 3 to 5). In reactions wherein the integration was mediated by the fusion protein IN/E2C (Fig. 5A, lanes 6 to 8) or E2C/IN (Fig. 5A, lanes 9 to 11), the E2C-binding site was not used as a target by the fusion proteins, and a significant fraction of the integration events instead occurred near the E2C-binding sequence. For both IN/E2C and E2C/IN, the integration hot spots were distributed asymmetrically and clustered within a 10-nucleotide region about 10 nucleotides upstream (5′) of the C-rich strand of the E2C-binding site. In comparison to the wild-type protein, there was a notable decrease in the frequency of integration in the outlying regions of the E2C-binding sequence. For IN/E2C, the decrease in nonspecific integration was seen primarily downstream of the E2C-binding site, whereas the nonspecific integration using E2C/IN was uniformly decreased throughout the target DNA molecule (Fig. 5A, lanes 6 to 8 versus 9 to 11).

Since the PCR-based assay only monitored the integration events occurring into the vicinity downstream of the reverse

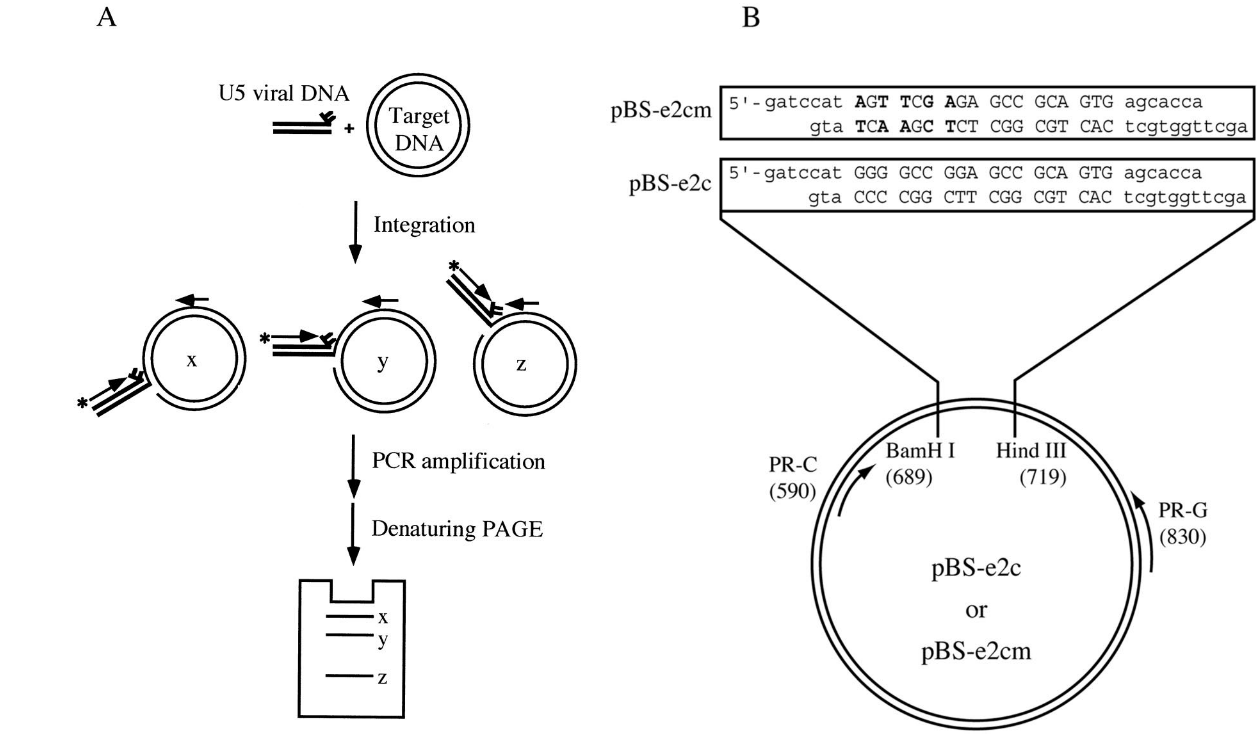
A
U5 viral DNA
+
Target DNA
Integration
x
y
z
*
PCR amplification
Denaturing PAGE
x
y
z
B
pBS-e2cm
5'-gatccat AGT TCG AGA GCC GCA GTG agcacca
gta TCA AGC TCT CGG CGT CAC tcgtggttcga
pBS-e2c
5'-gatccat GGG GCC GGA GCC GCA GTG agcacca
gta CCC CGG CTT CGG CGT CAC tcgtggttcga
PR-C (590)
BamH I (689)
Hind III (719)
PR-G (830)
pBS-e2c
or
pBS-e2cm

FIG. 4. PCR-based assay for determining distribution and preference of integration sites. (A) Schematic representation of the PCR-based assay. The reaction included the preprocessed U5 double-stranded oligonucleotide (thick lines) and supercoiled plasmid DNA (thin circular lines) as the donor and target substrates, respectively. Thick arrows denote the primers for the PCR. The 5′-end-labeled forward primer, indicated by the asterisk, annealed to the viral U5 DNA, while the reverse primer annealed to the target DNA. Integration of U5 viral DNA at different positions (denoted by x, y, and z) generated a population of recombinants products. The distribution of integration sites was analyzed by the lengths of the PCR products after separation on a denaturing polyacrylamide gel. Integration events occurring into the two target DNA strands were monitored separately by using a reverse primer that annealed to the top or bottom strand. (B) Target DNA substrate. The wild-type or mutant E2C-binding sequence (uppercase letters) was cloned between the *Bam*HI and *Hin*dIII sites of a plasmid derived from pBluescript II KS(+), resulting in pBS-e2c and pBS-e2cm, respectively. The point mutations in the E2C-binding sequence are marked in bold. The arrows represent the primers PR-G and PR-C used in the PCR amplification of the integration products occurring in the plasmid DNA containing the G-rich and C-rich strands of the E2C-binding site, respectively. The numbers in parentheses denote the map positions of the sites for primer annealing and restriction enzyme cleavage.

primer, we do not know the percentage of total integration events directed to the integration hot spots. However, if the quantitation of integration specificity was based only on the integration events occurring within the 530-bp region (nucleotide lengths 72 to 603) around the E2C-binding site, we found that 65 and 73% of integration mediated by 0.5 μM IN/E2C and E2C/IN, respectively, took place within the 10-nucleotide hot spot region (nucleotide lengths 154 to 163). Using the identical criteria for quantitation, we calculated that integration mediated by the same concentration of wild-type integrase into the same area only constituted 5% of the total integration events.

To ensure that the integration hot spots observed with the fusion proteins were not experimental artifacts, the integration reaction was carried out in the presence of a fixed amount of wild-type HIV-1 integrase and various amounts of E2C/GST protein (Fig. 5A, lanes 12 to 14). Similar to reactions observed earlier with either IN/E2C or E2C/IN fusion proteins, very few integration events took place within the E2C-binding site in the presence of E2C/GST and wild-type integrase. However, in contrast to IN/E2C or E2C/IN, integration hot spots were not detected near the E2C-binding site. The levels of nonspecific integration in the outlying regions were also not noticeably altered. The data provide support that the integration pattern, as defined by both the distribution and frequency of integration events, of IN/E2C or E2C/IN results from two components working in *cis* as a fusion protein and not from a combined effect of two separate functions provided in *trans* by individual components. The result also ruled out the possibility that the directed integration by the fusion proteins could be an indirect consequence of DNA distortion induced by protein binding of the E2C recognition site (24).

The distribution and frequency of integration events occurring into the target DNA strand containing the G-rich sequence of the E2C-binding site were also examined (Fig. 5B). Overall, integration hot spots were also observed on the G-rich strand, but they were less pronounced and more scattered than those on the C-rich strand. For IN/E2C, a major hot spot was found immediately upstream of the E2C-binding site, and several other hot spots were located within a 20-bp region downstream of the E2C-binding site (Fig. 5B, lanes 4 and 5). For E2C/IN, several hot spots were also seen within the 20-nucleotide region downstream of the E2C-binding site, but the major hot spot was located within the E2C-binding site (Fig. 5B,

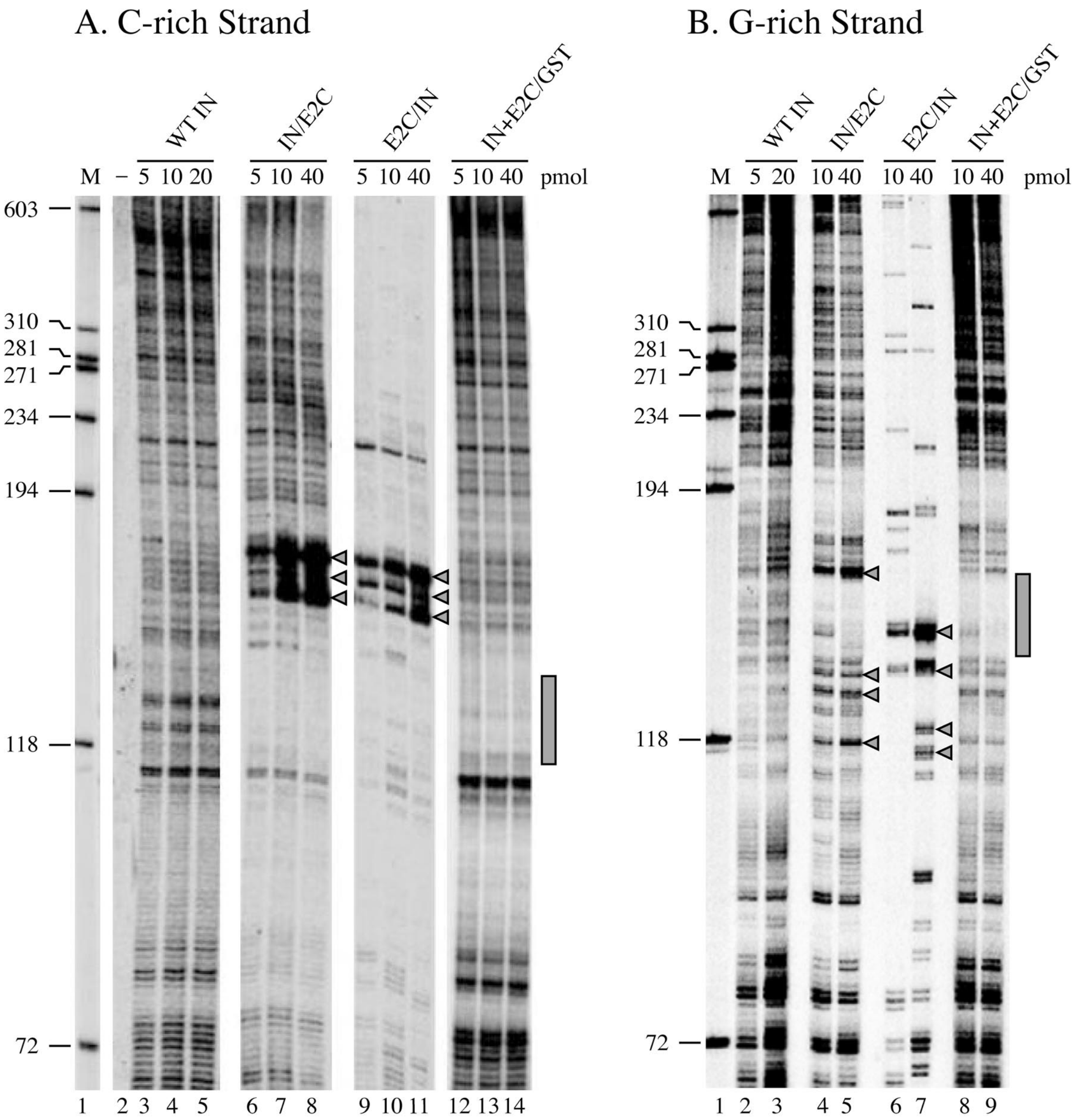

FIG. 5. Selection of target sites by wild-type HIV-1 integrase or fusion proteins containing HIV-1 integrase and E2C. (A) Distribution of integration sites on the C-rich strand of the E2C-binding site. One microgram of target DNA, pBS-e2c, was preincubated with the indicated concentrations of proteins (lanes 3 to 11) for 15 min at room temperature. The integration reaction was started by adding 0.3 pmol of preprocessed U5 DNA and incubating the mixture at 37°C for 45 min. The reaction products were amplified by PCR using labeled B2-1 and PR-C as forward and reverse primers, respectively. In lanes 12 to 14, pBS-e2c was preincubated with the E2C/GST protein (5, 10, or 40 pmol) for 10 min at room temperature. This was followed by the addition of 5 pmol of HIV-1 integrase and an additional preincubation period of 5 min at room temperature before the start of the integration reaction. Lane 1 contains the size marker, with the DNA lengths in nucleotides indicated on the left. Lane 2 is a negative control and represents the PCR products of an integration reaction carried out in the absence of enzymes. The stippled box on the right indicates the position of the E2C-binding site. Arrowheads denote the integration hot spots specific for E2C-containing fusion proteins. (B) Distribution of integration sites on the G-rich strand of the E2C-binding site. Experiments were performed identically to those described for panel A, except that PR-G was used as the reverse primer during PCR amplification. Other symbols have the same significance as in panel A.

lanes 6 and 7). Using the same method described earlier for quantitating integration specificity, we estimated that IN/E2C- and E2C/IN-mediated integration at the hot spots constituted 14 and 32%, respectively, of the total integration events, whereas 5% of integration mediated by wild-type integrase occurred into the same areas. Similar to the C-rich strand, integration hot spots were not detected when integrase and E2C/GST were added to the reaction as separate, individual proteins (Fig. 5B, lanes 8 and 9).

The activity of many integrase variants, although too weak to be detected by the oligonucleotide-based assays, can be studied using the more sensitive PCR-based assay (1, 20, 51). A previous study using integrase-LexA fusion protein showed that the ability to direct integration into a specific site can be

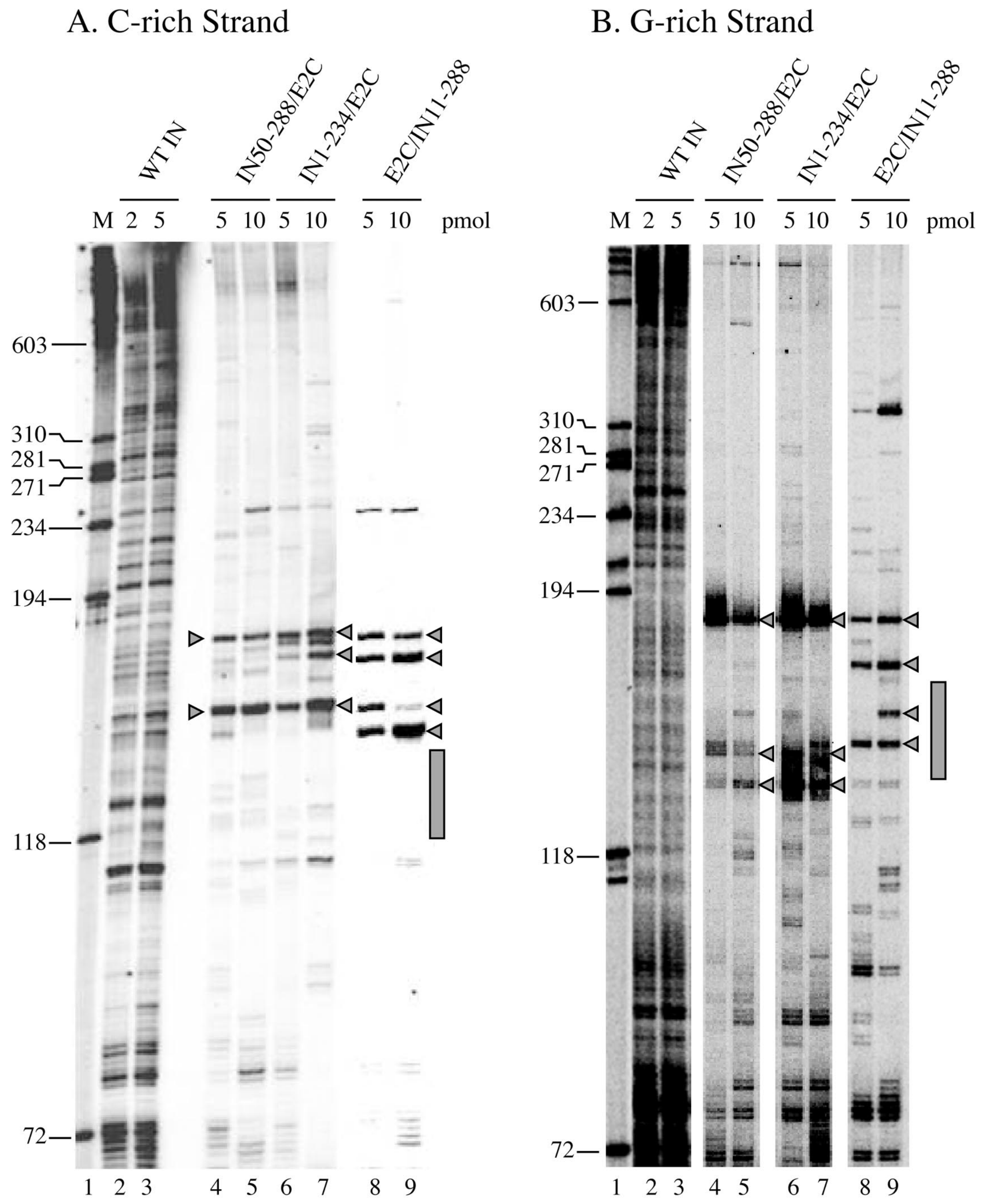

FIG. 6. Integration site usage of E2C fusion proteins containing N- or C-terminal-truncated HIV-1 integrase. (A) C-rich strand; (B) G-rich strand. Integration reactions were performed using wild-type HIV-1 integrase (lanes 2 and 3) or various fusion proteins containing N-terminal-truncated (lanes 4 and 5, IN50-288/E2C; lanes 8 and 9, E2C/IN11-288) or C-terminal-truncated integrase (lanes 6 and 7). Lane 1 contains the size marker, and the DNA lengths in nucleotides are indicated on the left. Symbols have the same significance as in Fig. 5.

achieved with a fusion protein containing the core domain (residues 50 to 234) of HIV-1 integrase (20). Three truncation variants of HIV-1 integrase fused to E2C were examined for their ability to mediate site-directed integration (Fig. 6). As expected from their low activities in oligonucleotide-based assays, the integration efficiencies of these truncated fusion proteins were poorer than that of their full-length counterpart. Other than the poor efficiency and minor differences in the specific choice of integration sites, the truncated fusion proteins IN50-288/E2C (Fig. 6A and B, lanes 4 and 5), IN1-234/E2C (Fig. 6A and B, lanes 6 and 7), and E2C/IN11-288 (Fig. 6A and B, lanes 8 and 9) showed integration patterns similar to those of the E2C fusion proteins containing a full-length integrase (Fig. 5A and B, lanes 6 to 11). For the strand containing the C-rich sequence of the E2C-binding site (Fig. 6A), the integration hot spots were localized within a 20-nucleotide region upstream of the E2C-binding site and integration within the E2C-binding site was absent. For the strand containing the G-rich sequence of the E2C-binding site (Fig. 6B), integration hot spots were present upstream and downstream, as well as within the E2C-binding site.

Site-directed integration by the fusion proteins depends on the correct recognition of the E2C-binding site. We interpreted that the relative absence of integration events within the E2C-binding site as well as the integration hot spots nearby were due to the enhanced presence of integrase-E2C fusion proteins at the E2C-binding site. To confirm that the directed integration by the fusion proteins relied on the presence of the E2C-binding sequence, integration reactions were carried out in the presence of pBS-e2cm DNA, which contains a mutant E2C-binding site (Fig. 4B). The binding affinity of E2C to the wild-type and mutant sequences differs by approximately 100-fold (5). With pBS-e2cm as the target DNA, the distribution and frequency of integration events into the C-rich strand by various integrase-E2C fusion proteins (Fig. 7, lanes 6 to 17) were indistinguishable from those of the wild-type integrase (Fig. 7, lanes 3 to 5). We found that, with either wild-type integrase or fusion proteins, integration events were not precluded from occurring into the mutant E2C-binding site, and integration hot spots unique to the fusion proteins were not detected near the mutant binding sequence. Hot spots and cold spots of integration by various fusion proteins were also not observed on the G-rich strand (data not shown). In addition to pBS-e2cm, similar results were obtained when two related plasmid DNAs with no E2C-binding sequence, pBluescript II KS(+) and pBS-LA (20), were used as the target DNA (data not shown). The data indicated that site-directed integration mediated by the fusion proteins depends on the proper recognition of the E2C-binding site and can be blocked by mutating the binding sequence.

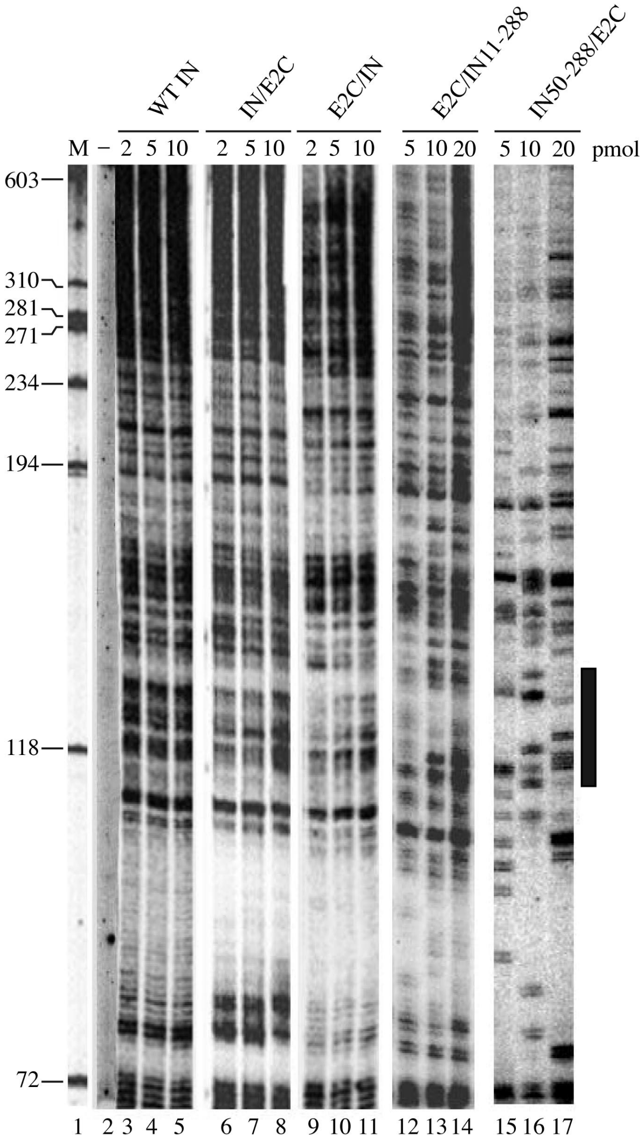

FIG. 7. Integration site selection of wild-type HIV-1 integrase and various fusion proteins in the presence of a mutant E2C-binding site. The integration reaction was performed in the presence of wild-type integrase (lanes 3 to 5) or the indicated fusion proteins (lanes 6 to 17) and 1 μg of pBS-e2cm, which contains a mutant E2C-binding site, as the target DNA. The integration products into the target DNA strand containing the C-rich sequence of the E2C-binding site were amplified using labeled B2-1 and PR-C as the forward and reverse primers, respectively. Lane 1 contains the DNA size marker, and lane 2 is an integration reaction carried out in the absence of enzymes. The filled box on the right indicates the position of the mutant E2C-binding site.

DISCUSSION

The major finding from these experiments is that integration of retroviral DNA in vitro can be directed to a predetermined site by using a fusion protein consisting of HIV-1 integrase and a polydactyl zinc finger protein, E2C. E2C can be tethered to the N or C terminus of HIV-1 integrase, and the resulting fusion proteins have similar DNA-binding specificities, catalytic activities, and integration site selectivities. Various deletion derivatives of HIV-1 integrase can also mediate site-directed integration when fused with E2C; however, their integration activities are significantly lower than that of the fusion protein containing the full-length wild-type integrase. Regardless of whether the fusion protein contains a wild-type or truncated integrase component, the ability to mediate site-directed integration depends on the presence of the cognate binding sequence of E2C on the target DNA.

Although site-directed integration has been reported previously with fusion proteins consisting of full-length or truncated retroviral integrase and various sequence-specific binding proteins (7, 8, 20, 29), the use of a polydactyl zinc finger protein as the target-specifying component offers important advantages over the published ones with regard to specificity and versatility. E2C belongs to a class of synthetic DNA-binding proteins constructed by linking two zinc finger proteins, with each containing three finger domains (4, 5, 38). Like the human and murine transcription factors Sp1 and Zif268, in which each zinc finger domain recognizes three nucleotides, the synthetic polydactyl zinc finger proteins specifically bind to an 18-bp contiguous DNA sequence (4, 5). Assuming random base distribu-

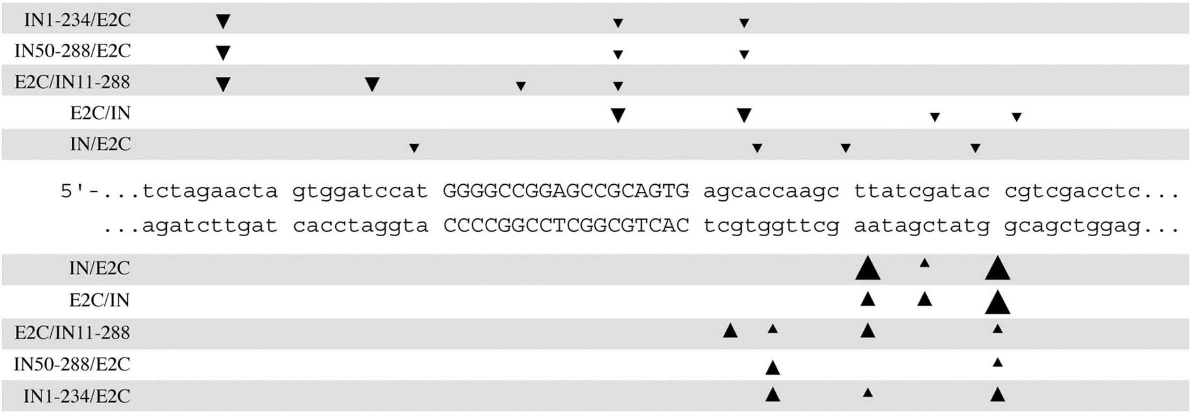
IN1-234/E2C
IN50-288/E2C
E2C/IN11-288
E2C/IN
IN/E2C
5'-...tctagaacta gtggatccat GGGGCCGGAGCCGCAGTG agcaccaagc ttatcgatac cgtcgacctc...
...agatcttgat cacctaggta CCCCGGCCTCGGCGTCAC tcgtggttcg aatagctatg gcagctggag...
IN/E2C
E2C/IN
E2C/IN11-288
IN50-288/E2C
IN1-234/E2C

FIG. 8. Positions of preferred integration sites of various HIV-1 integrase-E2C fusion proteins. The DNA sequence flanking the E2C-binding site of pBS-e2c is shown. The E2C recognition sequence is in uppercase letters. Arrowheads above and below the DNA sequence indicate the positions of preferred integration sites on the G-rich and C-rich strands of the E2C-binding site, respectively. The relative preferences of the integration sites for each fusion protein were determined by PhosphorImager analysis and are approximated using arrowheads of different sizes.

tion, an 18-bp address would be specific within 69 billion bp of sequence, more than sufficient for specifying a unique site in human and other mammalian genomes. For instance, a BLAST search of the GenBank database (human November 2002 freeze) verified that the 18-bp E2C-binding site is unique in the human genome and occurs only once on chromosome 17. Also, each zinc finger domain can potentially be developed as a modular building block for specific recognition of each of the 64 possible 5′-NNN-3′ sequences (5, 11, 16, 17, 31, 50). Depending on the sequence of the desired site, the modules can be assembled in any order necessary to form new six-zinc-finger proteins with specific recognition of that particular site. In addition to specificity and versatility, because the synthetic polydactyl zinc finger proteins are put through multiple rounds of selection for their target sequence, their binding affinities are typically in the subnanomolar to picomolar range, which are 10- to 100-fold higher than their three-zinc-finger counterparts and most other sequence-specific DNA-binding proteins (4, 5, 16, 31). The application of this class of designed DNA-binding proteins is also illustrated by studies in which artificial transcription factors based on modified zinc finger domains are used to activate or repress expression of reporter genes, as well as endogenous genes in the native chromosomal environment of animal and plant cells (4, 5, 10, 12, 16, 21, 28, 32, 33, 37, 47).

By examining the distribution and frequency of integration events on the target DNA, we found that the site-directed integration of viral DNA mediated by the integrase-E2C protein has similar characteristics to those reported previously (7, 8, 20, 29). The recognition sequence of the fusion protein is largely devoid of integration events, while integration hot spots specific for the fusion protein are located within 20 bp flanking the recognition sequence. Concomitantly, the frequency of integration events in the outlying regions (>20 bp) is decreased. These characteristics are consistent with our working model, in which the fusion protein binds to its cognate recognition site and mediates integration of viral DNA into the nearby regions. The absence of integration events in the cognate binding site is presumably a result of steric hindrance produced by the sequence-specific binding of the fusion protein. Retention of the fusion protein at the binding site in turn decreases the ability of the fusion protein to mediate integration events elsewhere on the target DNA molecule.

Although a majority of integration events mediated by the full-length integrase-E2C fusion proteins occur within a 20-bp region flanking the E2C recognition sequence, a considerable number of integration events are observed in the outlying region (20 bp or more) of the E2C-binding site. This is likely the result of the nonspecific DNA-binding activity of HIV-1 integrase (45, 55, 57). The preparation of several fusion proteins consisting of various truncated integrases was designed to test whether a higher specificity could be achieved by using an integrase without the domains known to interact with target DNA (23, 45, 55, 57). In comparison to the full-length fusion proteins, the integration specificity of the truncated fusion proteins was not improved, while the integration efficiency was significantly decreased. The result is consistent with the previous finding using integrase-LexA fusion proteins (20) and suggests that a better understanding of integrase-target DNA interaction is needed for suppressing the nonspecific integration activity of HIV-1 integrase.

One distinguishable feature of the site-directed integration mediated by the integrase-E2C fusion proteins is the asymmetric distribution of the integration hot spots. Although hot spots are found on both strands of the DNA helix, the major preferred sites are clustered upstream of the C-rich strand of the E2C recognition sequence (Fig. 8). We do not think that the absence of integration hot spots downstream of the C-rich strand of the E2C-binding site is attributable to local DNA sequences, since the same region is used efficiently as integration sites by wild-type integrase. Structural analysis of the Zif268-DNA complex showed that the zinc finger protein binds in the major groove, and most of the amino acid-DNA contacts are made with the G-rich strand of the target sequence (18, 41). Since the E2C-binding site is nonpalindromic, the binding of the fusion protein to the target DNA is directional and may result in an asymmetric distribution of integration hot spots. We are perplexed, however, by the observation that the same hot spots are found upstream of the C-rich strand regardless of whether the integrase is tethered to the N or C terminus of E2C. In the absence of structural information on the fusion protein, we do not know how the conformation and the position of the fusion protein in relation to the target DNA may affect the distribution of integration events.

In addition to asymmetric distribution, a majority of the integration hot spots mediated by each fusion protein have a ~10-bp periodicity (Fig. 8), suggesting that the fusion protein is bound at the E2C-binding site and interacts with the same face of the double helix. Certain fusion proteins, such as E2C/IN and E2C/IN11-288, have integration hot spots within the E2C-binding site, which is well protected against DNase I digestion by the footprinting analysis. Integration mediated by preintegration complexes containing the HIV-1 integrase-Zif268 fusion protein also shows hot spots within the Zif recognition site (8). It is possible that certain positions within the fusion protein-target DNA complex are more exposed and allow accessibility for the tethered integrase-donor DNA complex, but not for DNase I.

The ability of retroviruses to precisely and permanently introduce foreign genes into cellular chromosomes has resulted in their common use as vectors for both genetic engineering in higher eukaryotes and gene therapy. Because of their ability to infect nondividing cells, there is a strong interest in developing lentivirus-based vectors for gene delivery (40, 43). However, integration of viral DNA nonspecifically into host chromosomes is a major concern in the use of lentiviral and other retroviral vectors (22, 36, 53, 54). Studies on site-directed integration using fusion proteins may lead to a new approach for inserting exogenous genes at specific sites and improve the therapeutic application of current retroviral vectors.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants CA68859 to S.A.C. and GM065059 to D.J.S. and C.F.B.

We thank Michelle Holmes-Son and other members of the Chow laboratory for helpful discussions, Luke Deltredici for graphic support, and Michael Carey for comments on and critical reading of the manuscript.

REFERENCES

1. **Appa, R. S., C. G. Shin, P. Lee, and S. A. Chow.** 2001. Role of the nonspecific DNA-binding region and alpha helices within the core domain of retroviral integrase in selecting target DNA sites for integration. J. Biol. Chem. **276:** 45848–45855.
2. **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.
3. **Beerli, R. R., and C. F. Barbas III.** 2002. Engineering polydactyl zinc-finger transcription factors. Nat. Biotechnol. **20:**135–141.
4. **Beerli, R. R., B. Dreier, and C. F. Barbas III.** 2000. Positive and negative regulation of endogenous genes by designed transcription factors. Proc. Natl. Acad. Sci. USA **97:**1495–1500.
5. **Beerli, R. R., D. J. Segal, B. Dreier, and C. F. Barbas III.** 1998. Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. Proc. Natl. Acad. Sci. USA **95:**14628–14633.
6. **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.
7. **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.
8. **Bushman, F. D., 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. **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.
10. **Choo, Y., A. Castellanos, B. Garcia-Hernandez, I. Sanchez-Garcia, and A. Klug.** 1997. Promoter-specific activation of gene expression directed by bacteriophage-selected zinc fingers. J. Mol. Biol. **273:**525–532.
11. **Choo, Y., and A. Klug.** 1994. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc. Natl. Acad. Sci. USA **91:**11163–11167.
12. **Choo, Y., I. Sanchez-Garcia, and A. Klug.** 1994. In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. Nature **372:**642–645.
13. **Chow, S. A.** 1997. In vitro assays for activities of retroviral integrase. Methods **12:**306–317.
14. **Craigie, R.** 1992. Hotspots and warm spots: integration specificity of retroelements. Trends Genet. **8:**187–190.
15. **Desjarlais, J. R., and J. M. Berg.** 1994. Length-encoded multiplex binding site determination: application to zinc finger proteins. Proc. Natl. Acad. Sci. USA **91:**11099–11103.
16. **Dreier, B., R. R. Beerli, D. J. Segal, J. D. Flippin, and C. F. Barbas III.** 2001. Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. J. Biol. Chem. **276:**29466–29478.
17. **Dreier, B., D. J. Segal, and C. F. Barbas III.** 2000. Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. J. Mol. Biol. **303:**489–502.
18. **Elrod-Erickson, M., M. A. Rould, L. Nekludova, and C. O. Pabo.** 1996. Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. Structure **4:**1171–1180.
19. **Engelman, A., A. B. Hickman, and R. Craigie.** 1994. The core and carboxy-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. J. Virol. **68:**5911–5917.
20. **Goulaouic, H., and S. A. Chow.** 1996. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and *Escherichia coli* LexA protein. J. Virol. **70:**37–46.
21. **Guan, X., J. Stege, M. Kim, Z. Dahmani, N. Fan, P. Heifetz, C. F. Barbas III, and S. P. Briggs.** 2002. Heritable endogenous gene regulation in plants with designed polydactyl zinc finger transcription factors. Proc. Natl. Acad. Sci. USA **99:**13296–13301.
22. **Hacein-Bey-Abina, S., C. von Kalle, M. Schmidt, F. Le Deist, N. Wulffraat, E. McIntyre, I. Radford, J. L. Villeval, C. C. Fraser, M. Cavazzana-Calvo, and A. Fischer.** 2003. A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. **348:**255–256.
23. **Heuer, T. S., and P. O. Brown.** 1997. Mapping features of HIV-1 integrase near selected sites on viral and target DNA molecules in an active enzyme-DNA complex by photo-cross-linking. Biochemistry **36:**10655–10665.
24. **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.
25. **Hughes, S. H., P. R. Shank, D. H. Spector, H. J. Kung, J. M. Bishop, H. E. Varmus, P. K. Vogt, and M. L. Breitman.** 1978. Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. Cell **15:**1397–1410.
26. **Jordan, S. R., and C. O. Pabo.** 1988. Structure of the lambda complex at 2.5 Å resolution: details of the repressor-operator interactions. Science **242:**893–899.
27. **Kahn, E., J. P. G. Mack, R. A. Katz, J. Kulkosky, and A. M. Skalka.** 1991. Retroviral integrase domains: DNA binding and recognition of LTR sequences. Nucleic Acids Res. **19:**851–860.
28. **Kang, J. S., and J. S. Kim.** 2000. Zinc finger proteins as designer transcription factors. J. Biol. Chem. **275:**8742–8748.
29. **Katz, R. A., G. Merkel, and A. M. Skalka.** 1996. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. Virology **217:**178–190.
30. **Kay, M. A., J. C. Glorioso, and L. Naldini.** 2001. Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. Nat. Med. **7:**33–40.
31. **Kim, J.-S., and C. O. Pabo.** 1998. Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. Proc. Natl. Acad. Sci. USA **95:**2812–2817.
32. **Kim, J. S., J. Kim, K. L. Cepek, P. A. Sharp, and C. O. Pabo.** 1997. Design of TATA box-binding protein/zinc finger fusions for targeted regulation of gene expression. Proc. Natl. Acad. Sci. USA **94:**3616–3620.
33. **Kim, J. S., and C. O. Pabo.** 1997. Transcriptional repression by zinc finger peptides. Exploring the potential for applications in gene therapy. J. Biol. Chem. **272:**29795–29800.
34. **Lee, M. S., G. P. Gippert, K. V. Soman, D. A. Case, and P. E. Wright.** 1989. Three-dimensional solution structure of a single zinc finger DNA-binding domain. Science **245:**635–637.
35. **Lewis, L. K., G. R. Harlow, L. A. Gregg-Jolly, and D. W. Mount.** 1994. Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli.* J. Mol. Biol. **241:**507–523.
36. **Li, Z., J. Dullmann, B. Schiedlmeier, M. Schmidt, C. von Kalle, J. Meyer, M. Forster, C. Stocking, A. Wahlers, O. Frank, W. Ostertag, K. Kuhlcke, H. G. Eckert, B. Fehse, and C. Baum.** 2002. Murine leukemia induced by retroviral gene marking. Science **296:**497.
37. **Liu, P. Q., E. J. Rebar, L. Zhang, Q. Liu, A. C. Jamieson, Y. Liang, H. Qi,**

P. X. Li, B. Chen, M. C. Mendel, X. Zhong, Y. L. Lee, S. P. Eisenberg, S. K. Spratt, C. C. Case, and A. P. Wolffe. 2001. Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. J. Biol. Chem. **276:**11323–11334.

38. **Liu, Q., D. J. Segal, J. B. Ghiara, and C. F. Barbas III.** 1997. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. Proc. Natl. Acad. Sci. USA **94:**5525–5530.
39. **Miller, J., A. D. McLachlan, and A. Klug.** 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J. **4:**1609–1614.
40. **Naldini, L., U. Blomer, P. Gallay, D. Ory, R. Mulligan, F. H. Gage, I. M. Verma, and D. Trono.** 1996. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science **272:**263–267.
41. **Pavletich, N. P., and C. O. Pabo.** 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. Science **252:**809–817.
42. **Pfeifer, A., and I. M. Verma.** 2001. Gene therapy: promises and problems. Annu. Rev. Genomics Hum. Genet. **2:**177–211.
43. **Poeschla, E. M., F. Wong-Staal, and D. J. Looney.** 1998. Efficient transduction of nondividing human cells by feline immunodeficiency virus lentiviral vectors. Nat. Med. **4:**354–357.
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. **Puras Lutzke, R. A., C. Vink, and R. H. A. Plasterk.** 1994. Characterization of the minimal DNA-binding domain of the HIV integrase protein. Nucleic Acids Res. **22:**4125–4131.
46. **Rebar, E. J., H. A. Greisman, and C. O. Pabo.** 1996. Phage display methods for selecting zinc finger proteins with novel DNA-binding specificities. Methods Enzymol. **267:**129–149.
47. **Rebar, E. J., Y. Huang, R. Hickey, A. K. Nath, D. Meoli, S. Nath, B. Chen, L. Xu, Y. Liang, A. C. Jamieson, L. Zhang, S. K. Spratt, C. C. Case, A. Wolffe, and F. J. Giordano.** 2002. Induction of angiogenesis in a mouse model using engineered transcription factors. Nat. Med. **8:**1427–1432.
48. **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.
49. **Segal, D. J., and C. F. Barbas III.** 2001. Custom DNA-binding proteins come of age: polydactyl zinc-finger proteins. Curr. Opin. Biotechnol. **12:**632–637.
50. **Segal, D. J., B. Dreier, R. R. Beerli, and C. F. Barbas III.** 1999. Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. Proc. Natl. Acad. Sci. USA **96:**2758–2763.
51. **Shibagaki, Y., and S. A. Chow.** 1997. Central core domain of retroviral integrase is responsible for target site selection. J. Biol. Chem. **272:**8361–8369.
52. **Shih, C.-C., J. P. Stoye, and J. M. Coffin.** 1988. Highly preferred targets for retrovirus integration. Cell **53:**531–537.
53. **Temin, H. M.** 1990. Safety considerations in somatic gene therapy of human disease with retrovirus vectors. Hum. Gene Ther. **1:**111–123.
54. **Verma, I. M., and N. Somia.** 1997. Gene therapy—promises, problems and prospects. Nature **389:**239–242.
55. **Vink, C., A. A. M. Oude Groeneger, and R. H. A. Plasterk.** 1993. Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type 1 integrase protein. Nucleic Acids Res. **21:**1419–1425.
56. **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.
57. **Woerner, A. M., and C. J. Marcus-Sekura.** 1993. Characterization of a DNA binding domain in the C terminus of HIV-1 integrase by deletion mutagenesis. Nucleic Acids Res. **21:**3507–3511.