VIROLOGY 222, 283–288 (1996)
ARTICLE NO. 0422

SHORT COMMUNICATION

Target-Sequence Preferences of HIV-1 Integration Complexes *in Vitro*

YEOU-CHERNG BOR,* MICHAEL D. MILLER, FREDERIC D. BUSHMAN,[1] and LESLIE E. ORGEL

*The Salk Institute for Biological Studies, Post Office Box 85800, San Diego, California 92186-5800; and *Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan*

Received March 13, 1996; accepted May 28, 1996

Integration of reverse transcribed retroviral cDNA is not restricted to particular host DNA sequences. However, the frequency of integration into a particular phosphodiester bond is influenced by the local sequence. Here we examine the target-sequence preferences of purified HIV integrase and viral nucleoprotein complexes (preintegration complexes) isolated from freshly infected cells. We find that the three-base sequence including the integration site is not the major factor determining the frequency of integration, since identical triplets embedded in different sequences are used with very different efficiencies. However, there is a statistically significant bias against integration upstream of a pyrimidine nucleotide. The target-sequence preferences of purified integrase and preintegration complexes are very different. Strong integration sites on opposite DNA strands occur in pairs separated by five residues when preintegration complexes are used but not with purified integrase. These studies highlight the difference between the two sources of HIV integration activity and may provide the basis for a simple assay for the correct assembly of viral nucleoprotein complexes.

Integration of reverse transcribed retroviral cDNA into a chromosome of the host is a necessary step in retroviral replication. Integration is dependent on a viral-encoded integrase protein (*7, 11, 26, 32*) and specific DNA sequences at each end of an unintegrated viral DNA (*9, 26, 31*). During integration *in vivo,* the blunt ends of the linear product of reverse transcription are first cleaved to remove two nucleotides from each 3′-end. The recessed 3′ ends are then joined to protruding 5′ ends of breaks made in the target DNA. For the case of HIV, the points of joining on the two DNA strands are staggered by five bases in the 5′ direction. This intermediate is then processed, probably by host DNA repair enzymes, to yield an integrated provirus (see *14* for a review).

In reconstituted reactions *in vitro,* purified integrases can cleave two nucleotides from the 3′ end of a DNA substrate that resembles the viral end and direct covalent integration of the cleaved products into a target DNA (*4, 6, 10, 19, 20, 33, 37*). *In vivo,* integrase proteins and the viral DNAs are found in large nucleoprotein complexes called preintegration complexes. Such complexes have been partially purified from cells infected with Moloney murine leukemia virus (MoMLV), Avian leukosis virus (ALV), or Human immunodeficiency virus (HIV) (*2, 13, 15, 23*). These preintegration complexes are capable of directing the covalent connection of the endogenously synthesized retroviral DNA into a target DNA added *in vitro.*

[1] To whom correspondence and reprint requests should be addressed.

Several previous studies have exploited the *in vitro* integration systems to characterize the factors influencing the choice of integration target sites. Sites in naked DNA are used with different efficiencies, though integration at most sites is detectable (*5, 10, 19, 22*). In two studies of integration by simple retroviruses *in vitro,* MoMLV and ALV, the integration sites selected by purified integrase were similar to the sites selected by preintegration complexes (*21, 30*).

Here we characterize the DNA sequence preferences of HIV-1 integration complexes in detail and compare integration by purified HIV-1 integrase and HIV-1 preintegration complexes. In assays using purified integrase, a short duplex oligonucleotide matching one end of the unintegrated viral DNA (LTR) was used as the integration donor DNA. In assays with preintegration complexes, endogenously synthesized HIV cDNA served as the donor.

An oligonucleotide containing all possible three-base DNA sequences (triplet target) was designed to serve as an integration target (Table 1). We chose three base sequences because these are the longest that can be completely characterized with reasonable effort. Thirty-two different trinucleotide sequences are encoded in the top strand and the remaining 32 are encoded in the bottom strand. Triplets are overlapped and embedded in this 34-base pair segment without duplication. The triplet target oligonucleotide was cloned into a plasmid DNA for use as an integration target. Integration into flanking DNA sequences was also analyzed for both purified integrase and preintegration complexes. Furthermore, inte-

0042-6822/96 $18.00

TABLE 1

Sequences of DNA Oligonucleotides Used in This Study

Viral cDNA end mimic

FB64	5' ACTGCTAGAGATTTTCCACACGGATCCTAGGC 3'
FB65-2	3'ACGATCTCTAAAAGGTGTGCCTAGGATCCG 5'

PCR primers (viral end)

FB66	5' GCCTAGGATCCGTGTGGAAAATC 3'
FB642 (U3)	5' TGTGAATTAGCCCTTCCA 3'
FB652 (U5)	5' TGTGGAAAATCTCTAGCA 3'

PCR primers (target DNA)

1201 (top strand)	5' AACAGCTATGACCATG 3'
1211 (bottom strand)	5' GTAAAACGACGGCCAGT 3'

Integration target containing all triplet sequences

64 Top	5' AGACTACGAAATCAACAGCACCCTCCGCCATAAG 3'
64 Bottom	3' TCTGATGCTTTAGTTGTCGTGGGAGGCGGTATTC 5'

Note. The 64 possible DNA triplet sequences in the all-triplet target (bottom) are marked with brackets.

gration into a second plasmid target was analyzed for purified integrase.

The integration products were analyzed by a PCR method (Fig. 1) (*3, 21, 30*). Following integration *in vitro,* products were deproteinized and used as templates for PCR amplification. PCR primers were selected so that one primer was complementary to a target DNA sequence, and the other was complementary to the LTR terminus. The target primer was ^{32}P-labeled on its 5′ end. PCR amplification of integration products generated a population of molecules that were then denatured and analyzed on a DNA sequencing-type gel. Each band on the gel corresponded to integration at a specific phosphodiester bond. The frequency of integration at a particular site was reflected in the intensity of the corresponding band on the final autoradiogram. PCR primers to either side of the cloned triplet target were used to assay integration into each target DNA strand.

We included an "EDTA control" in each series of experiments. EDTA is known to suppress integration by chelating essential metal ions. We confirmed that no PCR amplification products were found if EDTA was included in the integration reaction buffer. This shows that the bands we observed when EDTA was omitted are due to integration products and not to amplification from unreacted substrate DNAs.

Figures 2A and 2B present the results of the PCR assay of integration sites in the triplet target. Integration into the top and bottom strands is illustrated in A and phosphorimager quantitation of the data is illustrated in B. It can be seen that the intensities of different bands vary over at least a 10-fold range.

Table 2 summarizes the intensity of integration brought about by purified integrase at each triplet in the all-triplet target. Additional data on utilization of triplets in flanking sequences and in another plasmid is also included. The integration sites for each triplet lay between the first and second nucleotide (i.e., 5′-N*NN-3′, where * marks the integration site). Triplet sequences are marked as hot (>500 arbitrary units), medium (100–500 units), or cold (<100 units) for integration frequency. The data in Table 2 show that integration frequency is not strongly correlated with the triplet sequence. For example, the frequently occurring sequence AAT is hot in one case, medium in seven cases, and cold in three cases.

The results of the PCR assay of integration brought about by preintegration complexes are also illustrated in Fig. 2. Integration into the triplet target and additional flanking sequences was studied. The results of these experiments are summarized in Table 3. Sequences

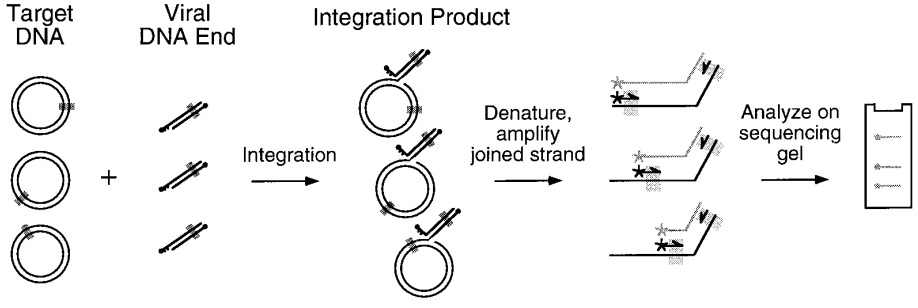

FIG. 1. PCR method for analyzing integration sites. Binding sites for primers used to amplify integration products are shown as grey rectangles. The diagram shows the method used for analysis of integration directed by purified integrase. A duplex oligonucleotide served as the donor DNA, and a circular plasmid as the target DNA. The method for analyzing integration directed by preintegration complexes is conceptually similar, but a 10-kb viral cDNA served as the integration donor and an isolated restriction fragment served as target. A plasmid containing all possible DNA sequence triplets (pLOUC, referred to here as "the triplet target") was constructed in two steps. Oligonucleotides 64 top and 64 bottom (Table 1) were annealed and ligated with pBS-SK+ that had been cleaved with *Eco*RV. Plasmids containing the desired insert (pLO) were isolated. DNA from this plasmid was found to give high backgrounds when used as an integration target in some PCR assays (described below), so the insert was recloned into the smaller plasmid pUC19. To accomplish this, a DNA fragment containing the cloned insert was isolated from pLO DNA by cleavage with *Eco*RI and *Hin*dIII and ligated with pUC19 DNA cleaved with *Eco*RI and *Hin*dIII, yielding plasmid pLOUC DNA.

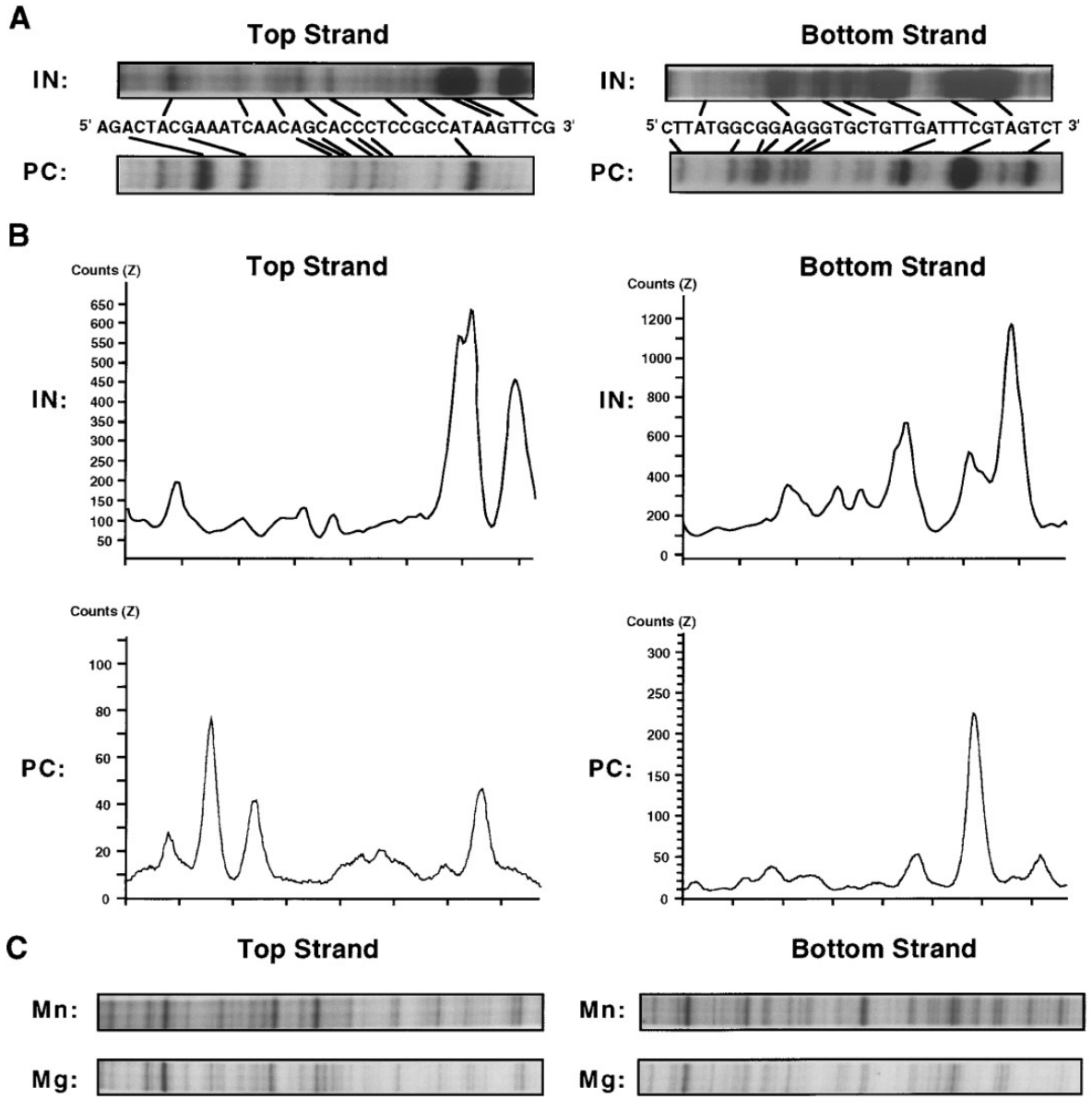

FIG. 2. Sites of integration used by purified HIV-1 integrase or preintegration complexes on the all triplet target. (A) Pattern of favored and disfavored sites for integration on the two DNA strands. For preintegration complexes, the U3 primer FB642 was used for the top strand and the U5 primer FB652 was used for the bottom strand. (B) Phosphorimager quantitation of the gel lanes in A. (C) Integration carried out by purified integrase in the presence of Mn^{2+} and Mg^{2+}. Integration was assayed by PCR into phage lambda target DNA as described (*3, 24*). In the integration assay using purified integrase, a duplex oligonucleotide, FB64/65-2 (Table 1), resembling the U5 viral end was used. 0.03 μM of FB64/65-2 was incubated with 0.26 μg of purified integrase (IN) for 10 min at 37° in integration buffer (5 m*M* $MnCl_2$, 20 m*M* HEPES (pH 7.0), 10 m*M* 2-mercaptoethanol, 10% (v/v) glycerol, 0.1 mg/ml bovine serum albumin), resulting in the stable assembly of integrase with the LTR DNA (*12*). The target DNA (final concentration 0.028 μM) was then added to the reaction to give a total volume of 20 μl. Reactions were incubated for a further 50 min at 37° and then stopped by adding 1 μl of 0.5 *M* EDTA and 1 μl of 0.1% SDS. The products were purified by phenolchloroform extraction and ethanol precipitation and then redissolved in 20 μl of TE. 2 μl of the integration product were used as template for the PCR amplification. Two plasmids were used as integration targets: pLOUC and pUC19(WT) (described in *1*). Preintegration complexes were prepared as described and partially purified by sucrose gradient centrifugation (*15, 24*). In the integration assay, 11 n*M* of DNA target (a *Pvu*II restriction fragment released from pLOUC containing the triplet target) was added to cytoplasmic extracts containing preintegration complexes in Buffer K. Reactions were incubated for 1 hr at 37°. The reaction was then stopped and the products purified as described (*24*). The integration products were dissolved in 20 μl of TE. 2 μl of this solution were used as template for PCR amplification. Integration products were analyzed by PCR essentially as described (*3, 21, 30*). Each integration product was assayed in two separate reactions containing either primer 1201 or 1211 from New England Biolabs as the target primer. Products of reactions with purified integrase were amplified with primer FB66 (viral end primer). Products of reactions with preintegration complexes were amplified with viral end primers FB652 (U5) or FB642 (U3) (Table 1). To visualize the amplification product, one primer was labeled. Control experiments revealed that similar results were obtained whether the target primer or the LTR primer was labeled. The molecular lengths of PCR products were determined by electrophoresis in lanes adjacent to size markers generated by Sanger DNA sequencing reactions. The hotspots in Tables 2 and 3 were catalogued by visual inspection, guided by Phosphorimager traces (quantitation based on Phosphorimager traces alone at times conflicted with the relative strengths of signals as judged by autoradiography due to inadequate resolution of weak bands adjacent to strong ones by the Phosphorimager).

TABLE 2

Analysis of Target Sites Used by Purified Integrase in the All-Triplet Target and Another Target

	H	M	C		H	M	C		H	M	C		H	M	C
TTT		3	3	TCT		1	1	TAT		1	1	TGT	1	1	
TTC		3	1	TCC			1	TAC		2		TGC		6	
TTA		4	2	TCA		4	1	TAA	2	4		TGA	1	2	
TTG	1	6	1	TCG		3		TAG	1			TGG		1	1
CTT		5	2	CTT		1	1	CAT		1	1	CGT		3	
CTC			1	CCC		1		CAC	1	1	1	CGC		1	1
CTA		1	1	CCA		3		CAA		7	1	CGA		4	1
CTG		1		CCG		1		CAG		2		CGG		1	1
ATT		4	2	ACT			1	AAT	1	7	3	AGT	1	1	
ATC		1	2	ACC			1	AAC	1	2	1	AGC		5	1
ATA		2		ACA		1	1	AAA	1	6	1	AGA	1	1	
ATG		1	1	ACG		5		AAG	1	4	1	AGG		1	
GTT		1	3	GCT		5		GAT	1	1	1	GGT			1
GTC			1	GCC		3		GAC		1	1	GGC		3	1
GTA		2	1	GCA		4	1	GAA	1	2	2	GGA		1	
GTG		2		GCG		2	1	GAG		1		GGG		1	

Note. H, hot; M, medium; C, cold. Integration at a specific triplet as described here indicates covalent joining 5′ of the second nucleotide in the triplet sequence.

were scored hot (>25 units), medium (10 to 25 units), or cold (<10 units). As we found for purified integrase, the efficiency of integration into triplet sequences was strongly dependent on the context of the triplet, since no clear relationship between triplet sequence and integration frequency was found. Comparison of the integration pattern generated by purified integrase to that generated by preintegration complexes reveals that the two are quite different.

Although the patterns of integration were different for the two types of integration complexes, integration by both was less frequent upstream of pyrimidine residues. This can be seen by comparing the left two columns with the right two columns in Tables 2 and 3. The bias was found to be significant for both tables (P values of <0.001 for Table 2 and <0.0188 for Table 3 by Fisher's Exact test).

Several further experiments were carried out to exclude possible artifacts. Multiple aliquots from the same integration reaction were amplified separately and

TABLE 3

Analysis of Target Sites Used by Partially Purified Preintegration Complexes in the All-Triplet Target and Flanking Sequences

	H	M	C		H	M	C		H	M	C		H	M	C
TTT			1	TCT			1	TAT			1	TGT		1	
TTC			3	TCC			1	TAC			1	TGC		1	
TTA		1	1	TCA			2	TAA		2		TGA		1	
TTG			2	TCG		1	1	TAG		1	2	TGG			2
CTT			2	CTT		1		CAT	1			CGT	2		
CTC			1	CCC		1		CAC		1		CGC		1	
CTA			2	CCA			1	CAA			2	CGA		2	1
CTG		1		CCG			1	CAG			1	CGG		1	
ATT			2	ACT	1	1		AAT			3	AGT		1	1
ATC			2	ACC		1		AAC		2	2	AGC			2
ATA		1	1	ACA			1	AAA		1		AGA		2	
ATG			1	ACG			2	AAG		1	1	AGG		1	
GTT		2	1	GCT			2	GAT	1		1	GGT			1
GTC	1			GCC			1	GAC	1			GGC		1	1
GTA		1		GCA		1	1	GAA	1	2	1	GGA		1	
GTG			1	GCG		1	1	GAG		1		GGG		1	

Note. Markings are the same as in Table 2.

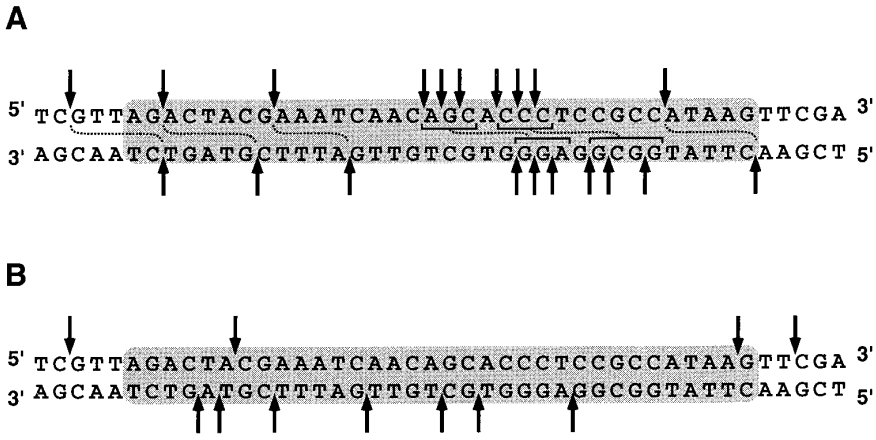

FIG. 3. Prominent sites of integration mapped on the triplet target sequence. (A) Sites used by preintegration complexes. The dotted lines indicate the linkage between integration sites on each DNA strand. (B) Sites used by purified integrase. Arrows mark the locations of prominent hotspots. The grey rectangles mark the all triplet target sequence.

shown to yield very similar gel patterns (data not shown). Thus the presence of hotspots and cold spots could not be a statistical artifact brought about by the amplification of very small numbers of integration products. Integration patterns for preintegration complexes in crude extracts were compared with those for complexes that had been partially purified by sucrose gradient sedimentation. The patterns generated were found to be closely similar, indicating that partial purification of preintegration complexes did not influence site selection (data not shown).

The unintegrated HIV cDNA contains two different terminal sequences, U3 and U5. Similar patterns of integration sites were seen regardless of whether the U5-specific or U3-specific primer was used for the analysis (data not shown). This finding confirms that the sequences of the primers used for amplification do not influence the resulting pattern of amplification products. This experiment also shows that the two cDNA ends in the preintegration complex are functionally equivalent with respect to target DNA recognition.

During normal integration of HIV *in vivo* the ligation junctions of the two viral DNA ends are inferred to be separated by 5 bp. Integration at a particular point on one strand should therefore be correlated with integration into the other strand at a separation of 5 bp. Characterization of the products of reactions *in vitro* with HIV-1 preintegration complexes confirms this expectation (*13*) (C. Farnet, personal communication). Figure 3A presents a map of prominent integration sites used by preintegration complexes in the triplet target. Most hotspots are matched by hotspots in the expected position 5 base pairs away on the opposite strand. For unknown reasons one site displayed a 4-bp spacing and another a 6-bp spacing. Our findings confirm that preintegration complexes *in vitro* direct coupled integration by the two ends of the viral cDNA. No similar correlation of integration sites on the two strands was observed in experiments using purified integrase (Fig. 3B).

We used Mg^{2+} in reactions containing preintegration complexes and Mn^{2+} in reactions with HIV-1 integrase. We do not believe this substitution is significant because purified integrase shows similar target-site preference with the two metal ions. Figure 3C presents a comparison of integration into a well studied region of the lambda DNA chromosome carried out by purified HIV-1 integrase in the presence of Mn^{2+} or Mg^{2+}. While slight differences can be seen, the overall pattern of hot and cold spots is generally similar.

Information on integration site preferences is available for three other integration systems, those of MoMLV, ALV, and the yeast retrotransposon Ty1. Retrotransposition by the Ty elements is carried out by reverse transcriptase and integrase enzymes similar in function and sequence to those of retroviruses. Ty1 integration is disfavored 5′ of T residues (*18*). MoMLV integration is favored 3′ of the sequence 5′-TpN-3′ (*28*), but no bias of the base 5′ of the point of joining was reported. ALV integration is favored 5′ of G or C residues (*16*). Here we report that HIV integration is favored 5′ of a purine residue. Sequences of integration sites used by HIV-1 *in vivo,* although few in number, also display more frequent joining 5′ of a purine residue (*34–36*) (S. Carteau and F. D. B., unpublished data). Evidently HIV, MoMLV, ALV, and Ty1, each have distinct target sequence preferences.

Previous comparisons of integration by purified integrase and by preintegration complexes from simple retroviruses have established that the two sources of integration activity exhibit similar target-site preferences. This was true for integration by MoMLV into naked DNA, protein DNA complexes, and minichromosomes. In each case only minor differences in the patterns of integration were detected (*29, 30*). In the case of ALV, purified integrase and preintegration complexes again showed similar but not identical target-site preferences (*21*). In our experiments the target-site preferences of purified integrase and preintegration complexes of HIV are essentially uncorrelated. In this respect, therefore, HIV differs fundamentally from MoMLV and ALV.

HIV preintegration complexes carry out coordinated integration of the two ends of the viral DNA (*13, 15*), while purified integrase incorporates viral ends (almost) independently of each other (*5, 6*). Perhaps the constraints needed to hold the viral ends together during integration influence target site selection. It may be relevant that purified integrases from ALV and MoMLV are more efficient than purified HIV integrase in bringing about coupled insertions of viral ends (*4, 17, 19*). Perhaps, for these integrases, the conformations of the complexes with purified integrase and preintegration complexes are more similar than they are for HIV. This might help to explain the surprising differences in the integration-site preferences of purified HIV integrase and preintegration complexes reported above. The method for detection of 5-bp staggers in HIV integration sites reported here may serve as a useful assay in future studies for the correct assembly of preintegration complexes from purified components.

ACKNOWLEDGMENTS

This work was supported by Grants AI34786 and AI37489-02 (F.D.B.) and AI29850-05 (L.E.O.) from the National Institutes of Health. F.D.B. is a Scholar of the Leukemia Society of America. We thank Verna Stitt for artwork, Sylvia Bailey for manuscript preparation, and members of the Bushman laboratory for comments on the manuscript.

REFERENCES

1. Bor, Y.-c., Bushman, F. D., and Orgel, L. E., *Proc. Natl. Acad. Sci. USA* 92, 10334–10338 (1995).
2. Brown, P. O., Bowerman, B., Varmus, H. E., and Bishop, J. M., *Cell* 49, 347–356 (1987).
3. Bushman, F. D., *Proc. Natl. Acad. Sci. USA* 91, 9233–9237 (1994).
4. Bushman, F. D., and Craigie, R., *J. Virol.* 64, 5645–5648 (1990).
5. Bushman, F. D., and Craigie, R., *Proc. Natl. Acad. Sci. USA* 88, 1339–1343 (1991).
6. Bushman, F. D., Fujiwara, T., and Craigie, R., *Science* 249, 1555–1558 (1990).
7. Clavel, F., Hoggan, M. D., Willey, R. L., Strebel, K., Martin, M., and Repaske, R., *J. Virol.* 63, 1455–1459 (1989).
8. Coffin, J. M., *In* "Virology" (B. N. Fields and D. M. Kinpe, Eds.), 2nd ed., pp. 1437–1500. Raven Press, New York, 1989.
9. Colicelli, J., and Goff, S. P., *Cell* 42, 573–580 (1985).
10. Craigie, R., Fujiwara, T., and Bushman, F., *Cell* 62, 829–837 (1990).
11. Donehower, L. A., and Varmus, H. E., *Proc. Natl. Acad. Sci. USA* 81, 6461–6465 (1984).
12. Ellison, V., and Brown, P. O., *Proc. Natl. Acad. Sci. USA* 91, 7316–7320 (1994).
13. Ellison, V. H., Abrams, H., Roe, T., Lifson, J., and Brown, P. O., *J. Virol.* 64, 2711–2715 (1990).
14. Farnet, C. M., and Bushman, F. D., *AIDS* 10, S3–S11 (1996).
15. Farnet, C. M., and Haseltine, W. A., *Proc. Natl. Acad. Sci. USA* 87, 4164–4168 (1990).
16. Fitzgerald, M. L., and Grandgenett, D. P., *J. Virol.* 68, 4314–4321 (1994).
17. Fitzgerald, M. L., Vora, A. C., Zeh, W. G., and Grandgenett, D. P., *J. Virol.* 66, 6257–6263 (1992).
18. Ji, H., Moore, D. P., Blomberg, M. A., Braiterman, L. T., Voytas, D. F., Natsoulis, G., and Boeke, J. D., *Cell* 73, 1–20 (1993).
19. Katz, R. A., Merkel, G., Kulkosky, J., Leis, J., and Skalka, A. M., *Cell* 63, 87–95 (1990).
20. Katzman, M., Katz, R. A., Skalka, A. M., and Leis, J., *J. Virol.* 63, 5319–5327 (1989).
21. Kitamura, Y., Lee, Y. M., and Coffin, J. M., *Proc. Natl. Acad. Sci. USA* 89, 5532–5536 (1992).
22. Leavitt, A. D., Rose, R. B., and Varmus, H. E., *J. Virol.* 66, 2359–2368 (1992).
23. Lee, Y. M. H., and Coffin, J. M., *J. Virol.* 64, 5958–5965 (1990).
24. Miller, M., Wang, B., and Bushman, F. D., *J. Virol.* 69, 3938–3944 (1995).
25. Miller, M. D., Bor, Y.-c., and Bushman, F. D., *Curr. Biol.* 5, 1047–1056 (1995).
26. Panganiban, A. T., and Temin, H. M., *Nature* 306, 155–160 (1983).
27. Panganiban, A. T., and Temin, H. M., *Proc. Natl. Acad. Sci. USA* 81, 7885–7889 (1984).
28. Pryciak, P., Muller, H.-P., and Varmus, H. E., *Proc. Natl. Acad. Sci. USA* 89, 9237–9241 (1992).
29. Pryciak, P. M., Sil, A., and Varmus, H. E., *EMBO J.* 11, 291–303 (1992).
30. Pryciak, P. M., and Varmus, H. E., *Cell* 69, 769–780 (1992).
31. Reicin, A. S., Kalpana, G., Paik, S., Marmon, S., and Goff, S., *J. Virol.* 69, 5904–5907 (1995).
32. Schwartzberg, P., Colecilli, J., and Goff, S. P., *Cell* 37, 1043–1052 (1984).
33. Sherman, P. A., and Fyfe, J. A., *Proc. Natl. Acad. Sci. USA* 87, 5119–5123 (1990).
34. Stevens, S. W., and Griffith, J. D., *Proc. Natl. Acad. Sci. USA* 91, 5557–5561 (1994).
35. Vincent, K. A., Ellison, V., Chow, S. A., and Brown, P. O., *J. Virol.* 67, 425–437 (1993).
36. Vink, C., van Gent, D. C., and Plasterk, R. H., *J. Virol.* 64(10), 5219–5222 (1990).
37. Vora, A. C., Fitzgerald, M. L., and Grandgenett, D. P., *J. Virol.* 64(11), 5656–5659 (1990).