JOURNAL OF VIROLOGY, July 1994, p. 4314–4321
0022-538X/94/$04.00+0

Vol. 68, No. 7

Retroviral Integration: In Vitro Host Site Selection by Avian Integrase

MICHAEL L. FITZGERALD AND DUANE P. GRANDGENETT*

Institute for Molecular Virology, St. Louis University Medical Center, St. Louis, Missouri 63110

Received 24 January 1994/Accepted 1 April 1994

Viral integrase catalyzes the integration of the linear viral DNA genome into the chromatin of the infected host cell, an essential step in the life cycle of retroviruses. The reaction produces a characteristic small duplication of host sequences at the site of integration, implying that there is a close juxtaposition of the viral DNA ends during a concerted integration event. We have used an in vitro assay to measure the concerted integration of virus-like plasmid DNA into naked lambda DNA catalyzed by virion purified avian integrase. In contrast to in vivo avian integration, which has strong fidelity for a 6-bp duplication, purified avian integrase in the context of this assay produced a distribution of duplication sizes, with the 6-bp size dominating. The metal cofactor Mg^{2+} induced increased fidelity for the 6-bp duplication relative to that with Mn^{2+}. The immediate sequence of the host site may also influence duplication size in that we found sites that sustained multiple independent integration events producing the same duplication size. Additionally, for each set of cloned integration sites (5, 6, and 7 bp), a unique but similar symmetrical pattern of G/C and A/T sequence biases was found. Using duplex oligonucleotides as target substrates, we tested the significance of the 6-bp G/C and A/T pattern for site selection. In the context of this assay, which is likely dominated by the integration of only one viral end, the 6-bp pattern was not preferred. Instead, integration was predominately into the 3′ ends of the oligonucleotides. The combined results of the lambda and oligonucleotide assays indicated that although host site selection has properties in common with recognition of the viral DNA termini, the nonrandom sequence preferences seen for host site selection were not identical to the sequence requirements for long terminal repeat recognition.

Early events in the life cycle of retroviruses include entry into the host cell, reverse transcription of the viral RNA genome, and subsequent integration of the viral DNA genome into the chromatin of the infected cell. The process of integration is a two-step reaction catalyzed by viral integrase (IN). First, the blunt-ended long terminal repeat (LTR) termini of the linear DNA genome are cleaved by two nucleotides on the 3′ ends (LTR cleavage or trimming). Cleavage by IN produces the conserved CA-OH-3′ ends that are then ligated into the host DNA at the site of integration (strand transfer). Strand transfer is also catalyzed by IN, with both strand transfer and LTR cleavage appearing to occur by a direct nucleophilic mechanism (8, 13). During trimming, the blunt LTR DNA termini serve as the substrates for nucleophilic attack by water molecules, the assumed in vivo nucleophiles (8). In strand transfer the host DNA is the substrate for nucleophilic attack, while the trimmed LTR termini serve as the nucleophiles. We have previously shown that avian IN recognizes the LTR termini in a similar manner when these termini are serving as the substrates of nucleophilic attack during trimming (9, 26) and as nucleophiles during strand transfer (10). For recent reviews of integration and the biochemistry of the reaction, see references 14, 21, and 27.

A specific number of host bases are duplicated during concerted integration of both viral DNA termini (full-site reaction). The size of the short DNA duplication is controlled by the infecting virus. In vivo, fidelity to a specific duplication size appears to be relatively strong (24). However, with purified IN, fidelity to one size is reduced (4, 10, 17). It has been suggested that other components of the in vivo integration complex are responsible for the maintenance of fidelity (4). In this report, we show that the metal cofactor present and the sequence of naked target DNA may also influence fidelity to one duplication size.

An additional consideration with respect to integration is what role the primary sequence of the host site DNA has in site selection. Although retroviruses can integrate into many different host sites, these events appear to be nonrandom in vivo (23). Whether the nonrandom aspects of cellular integration are due to chromatin structure, transcriptional activity, or primary sequence has not been resolved. In vitro, when either cell extracts from virus-infected cells (19, 22) or purified avian myeloblastosis virus (AMV) IN (10, 19) is used for integration activity, the selection of target sites on naked DNA appears to be nonrandom. The AMV results indicate that G/C base pairs at the sites of strand transfer are preferred (10, 15, 19). Additionally, for naked host DNA, sequences flanking the duplicated base pairs appear to influence site selection for both murine leukemia virus (MLV) integration and AMV integration (10, 15, 22).

We have continued to investigate AMV IN target site selection by using a lambda genetic assay that measures the concerted integration of model virus-like plasmid DNA into naked lambda DNA (10). Further indications were found that integration site preference is likely nonrandom on naked DNA and that the immediate flanking DNA target sequence beyond the duplicated base pairs has influence on site selection. A symmetrical pattern of G/C and A/T biases at the sites of strand transfer and in positions flanking the duplicated site was evident. The symmetrical pattern of sequence biases appeared to be independent of the divalent metal ion used.

We directly tested whether this same target nucleotide symmetry could be preferentially recognized by AMV IN within the context of a target oligonucleotide (16, 20) with an LTR oligonucleotide as donor. The data showed that IN did not recognize the internalized symmetrical target site on the oligonucleotide but rather that the 3′ ends served as preferred

* Corresponding author. Mailing address: Institute for Molecular Virology, St. Louis University, 3681 Park Ave., St. Louis, MO 63110. Phone: (314) 577-8411. Fax: (314) 577-8406.

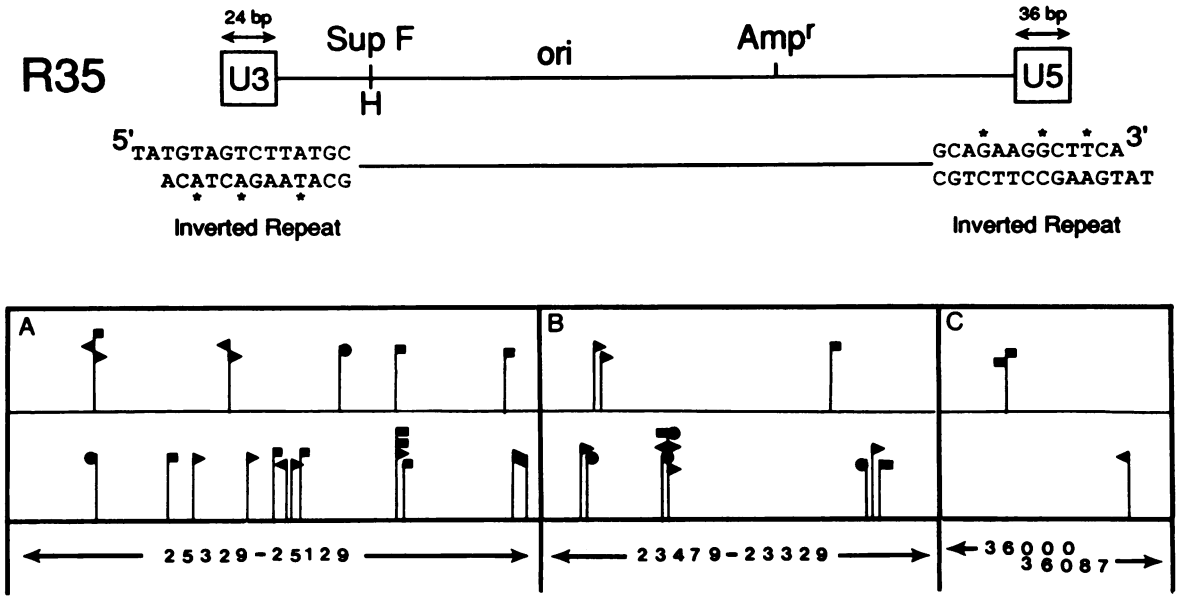

FIG. 1. R35 and selected multiple insertion sites. (Top) Structure of an *Nde*I-linearized LTR plasmid (R35). The plasmid is derived from pGEM-3 (Promega) and has 24 bp of avian U3 LTR sequence on one end and 36 bp of U5 LTR sequence on the other end (10). When juxtaposed, these ends form an imperfect inverted repeat (asterisks show nonsymmetrical positions). Two derivatives of R35, R33 and R55, were generated (not shown). R33 and R55 had either the 24 bp of U3 sequence at both ends or the 36 bp of U5 sequence at both ends, respectively. The plasmids carry the SupF gene (inserted at the *Hin*dIII site of pGEM-3) and the ampicillin resistance gene for selection in lambda gtWES and plasmid rescue (10). (Bottom) Isolated integration products were sequenced for analysis of site structure and location within the genome of lambda gtWES. Shown are selected regions demonstrating multiple integration sites. Symbols: ⚑, R35; ⚑, R33; ⚑, R55. The direction of the flag indicates the plasmid orientation within the lambda genome. The bottom tier of flags is for 6-bp duplications; the top tier in panels A and B is for 5-bp duplications, and the top tier in panel C is for 7-bp duplications. (A) A 5-bp preferred site, with three 5-bp and one 6-bp coincident duplications (three coincident 6-bp duplications were not analyzed). (B) A 6-bp hot spot, with five coincident duplications and one nearby. (C) The two coincident 7-bp duplications. Below the flag map are the sequence regions in lambda gtWES which are represented.

targets for this half-site strand transfer reaction (21). The preference for 3′ ends was not directly related to the use of LTR inverted repeat sequences as host sites.

MATERIALS AND METHODS

Concerted integration assay. Plasmids containing a small amount of viral terminal sequence can be linearized by the restriction enzyme *Nde*I to produce molecules that mimic the trimmed viral DNA genome (11). Such a plasmid (R35) is depicted at the top of Fig. 1 (see Fig. 1 legend for the LTR sequence structures of R35 and the derivative plasmids R33 and R55). The recessed ends obviate the cleavage step of integration. We used purified AMV IN to integrate R35 and derivative plasmids into naked lambda gtWES DNA (1). We have previously published details of the full-site reaction conditions and construction of the LTR plasmids (10). DNA from the reaction was purified and packaged to produce lambda phage. Selection for SupF-positive phage and cloning by plasmid rescue allowed for the analysis of the duplication structure and location of the integration sites (lambda gtWES can grow in a SupF$^-$ host only if an LTR plasmid has been integrated into its nonessential DNA). SupF-positive phage were amplified either as individual clones or as a mixed population (10). Phage DNA was isolated and cut with the restriction enzyme *Eco*RV. *Eco*RV digests lambda DNA but not the LTR plasmids. The *Eco*RV reaction mixture was diluted, ligated, and transformed into HB101 cells. Plasmids from Amp^r colonies were dideoxy sequenced with primers that anneal in the T7 and Sp6 regions of the pGEM-3-rescued LTR plasmids.

Statistical analysis. (i) G/C pattern significance. The significance of the G/C values of the 6- and 5-bp duplication sets was tested by determining whether the frequencies of G/C and A/T base pairs at each position were drawn from a homogeneous binomial distribution, by using a chi-square test (10).

(ii) Poisson distribution. A Poisson analysis was done on the distribution of the 6-bp sites cloned from the central nonessential region of lambda gtWES (see Fig. 4 of reference 10, region 21 to 24.7 kb). Deviation from a Poisson distribution indicates that the insertion events do not occur as a random Poisson process. To test this hypothesis, the nonessential region was subdivided into 148 25-bp intervals. The number of virus-like DNA recombinants in each region was determined. The observed integration frequency was compared with the expected frequency, assuming a random Poisson process, by using a chi-square test (see Table 2).

(iii) Preferred insertion site probability. Because multiple insertion sites were observed, a second analysis was conducted to assess the probability of these events. This was accomplished by first estimating the total number of possible integration sites. A value of 10,800 was derived from the assumption that 27% of the lambda gtWES genome can tolerate insertions (1). Dividing the number of cloned integration events by the number of possible integration sites gave the probability of a single insertion. The probability for multiple insertions at one site is derived by taking the probability of a single insertion to the power of the number of events at a multiple insertion site (see Table 3).

Oligonucleotide assay reaction conditions. Several oligonucleotide substrates (see Fig. 3) were synthesized (Operon, Inc.)

TABLE 1. Site type distribution[a]

Site type (bp)	% with: Mg^{2+} (n = 215)	% with: Mn^{2+} (n = 75)
Deletions		
>500	6.0	0
12–50[b]	1.9	10.7
Duplications		
4	0.5	9.3
5	14.9	24.0
6	64.2	44.0
7	7.9	5.3
8	0	4.0
11–50[b]	0.5	2.7
>50	4.2	0

[a] The integration products cloned had a variety of site types, with the avian-specific size dominating. The Mg^{2+} set (n = 215) had greater fidelity to the 6-bp size (64%) but also produced apparent duplications and deletions with sizes ranging up to 3,000 bp. With Mn^{2+} (n = 75) the distribution was shifted away from the 6-bp size, but no large duplications and deletions were found.
[b] A group of deletions or duplications whose sizes ranged as shown.

and labeled either at their 5′ ends with T4 polynucleotide kinase or at their 3′ ends by filling in with Klenow polymerase according to protocols of the manufacturer (Promega). Oligonucleotide assays were done by mixing the labeled DNA substrates on ice in a defined buffer (1% glycerol, 10 mM *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid [pH 7.5], 10 mM dithiothreitol, 50 mM NaCl, and either 5 mM $MgCl_2$ or 2 mM $MnCl_2$). IN was added, and the reaction mixture was incubated for 10 min on ice and then for 30 min at 37°C. The standard reaction mixture volume was 20 μl, with IN at 0.27 μM, 3′-labeled substrates at 0.04 μM, and 5′-labeled substrates at 0.06 μM. Unlabeled LTR donor DNA was added at 1.0 μM. Reactions were stopped by adding 81 μl of stop buffer (70 μl of water, 10 μl of 3 M sodium acetate, 1 μg of glycogen); this was followed by phenol-chloroform and ether extractions and ethanol precipitation. Reaction products were resuspended in denaturing buffer and were run on 12% sequencing gels. Autoradiography was overnight at −70°C with an intensifying screen.

RESULTS

Host duplication size and nonrandom integration. Previously, we had analyzed approximately 90 virus-like DNA recombinants inserted by AMV IN into lambda DNA (10). We wanted to further investigate the host duplication size fidelity of concerted integration events and to determine if the distribution of virus-like plasmids into the nonessential region was nonrandom. We also tested the effect of Mg^{2+} and Mn^{2+} on these integration parameters.

We have now sequenced nearly 300 integration products derived from the LTR plasmid/phage system and have found a distribution of site types, with the 6-bp avian-specific duplication predominating (Table 1). Interestingly, the metal cofactor supplied in the reaction mixture affected the distribution, with Mn^{2+} producing fewer 6-bp sites relative to those produced by Mg^{2+}. The smallest duplication observed was 4 bp, whereas the smallest deletion was 12 bp.

The integration products were mapped for their positions in the nonessential region of the lambda genome. A Poisson analysis of the 6-bp avian-specific duplicated sites showed the distribution to be nonrandom (Table 2). Sites which had sustained multiple independent insertion events (preferred integration sites) were found, and some of these sites appeared to be correlated with a specific duplication size (Fig. 1, bottom). The probability of occurrence for these events ranged from 1 in 4×10^6 for the 7-bp site to 1 in 1×10^9 for the 6-bp site (Table 3). In addition, the 6- and 5-bp sites were located in a region (3,800 bp, essentially the same as that used for the Poisson analysis) which could obviously tolerate insertions. Even with this reduced integration space, the events appeared to be nonrandom (Table 3).

TABLE 2. Poisson analysis for the distribution of 6-bp sites[a]

No. of recombinants/ interval (k)	No. of occurrences: Observed	No. of occurrences: Expected	P ($X = k$)	Contribution to chi-square
0	79	66.70	0.4507	2.27
1	37	53.16	0.3502	4.91
2	21	21.18	0.1431	0.00
3	7	5.63	0.0380	0.33
4	3	1.12	0.0076	3.16
5	0	0.18	0.0012	0.18
6	1	0.03	0.0002	31.36

[a] See Materials and Methods for the determination of the integration space. The total chi-square with 6 degrees of freedom is 42.03 ($P < 0.001$). The significance of the chi-square test indicates that the data deviated significantly from an expected random Poisson distribution.

It was possible that the observed biases were a result of lambda growth and plasmid rescue (see Materials and Methods). With this method, bias towards multiple insertions not related to the integration reaction could occur by at least two routes. First, it would be possible to contaminate cloned DNA from an integration reaction with previously cloned DNA. This possibility was ruled out by including for multiple integration sites only those events with different plasmid orientations and those events with the same orientation and plasmid type but sequenced from different experiments. That events with identical orientations and plasmid types were the results of independent integration events was also assured by the fact that the type of plasmid DNA (R35, R33, or R55) used in a reaction was consistently the type cloned out during the sequence analysis.

A second systematic cloning bias could result from selection during lambda phage growth or *Eco*RV plasmid rescue. These biases can be ruled out by two observations. First, lambda phage was grown either as individual clones (single plaque lifts after packaging) or as a population of clones (amplification of multiple plaques in one culture). If there was lambda growth selection, recombinants cloned by the population method would be biased towards preferred integration sites, and recombinants amplified as individual clones would not show the bias. This was not the case. Recombinants sequenced by the population method were divided evenly between preferred

TABLE 3. Probability of finding multiple insertion sites[a]

Duplication size (bp)	No. of recombinants	No. of possible positions	Probability of single insertion	No. of insertions	Probability of insertions at single site
6	171	10,800	0.016	5	$<1.0 \times 10^{-9}$
5	50	10,800	0.005	3	$<1.0 \times 10^{-7}$
7	22	10,800	0.002	2	4.1×10^{-6}
6	121	3,800	0.032	5	$<1.0 \times 10^{-7}$
5	41	3,800	0.011	3	1.3×10^{-6}

[a] See Materials and Methods for determination of possible integration positions.

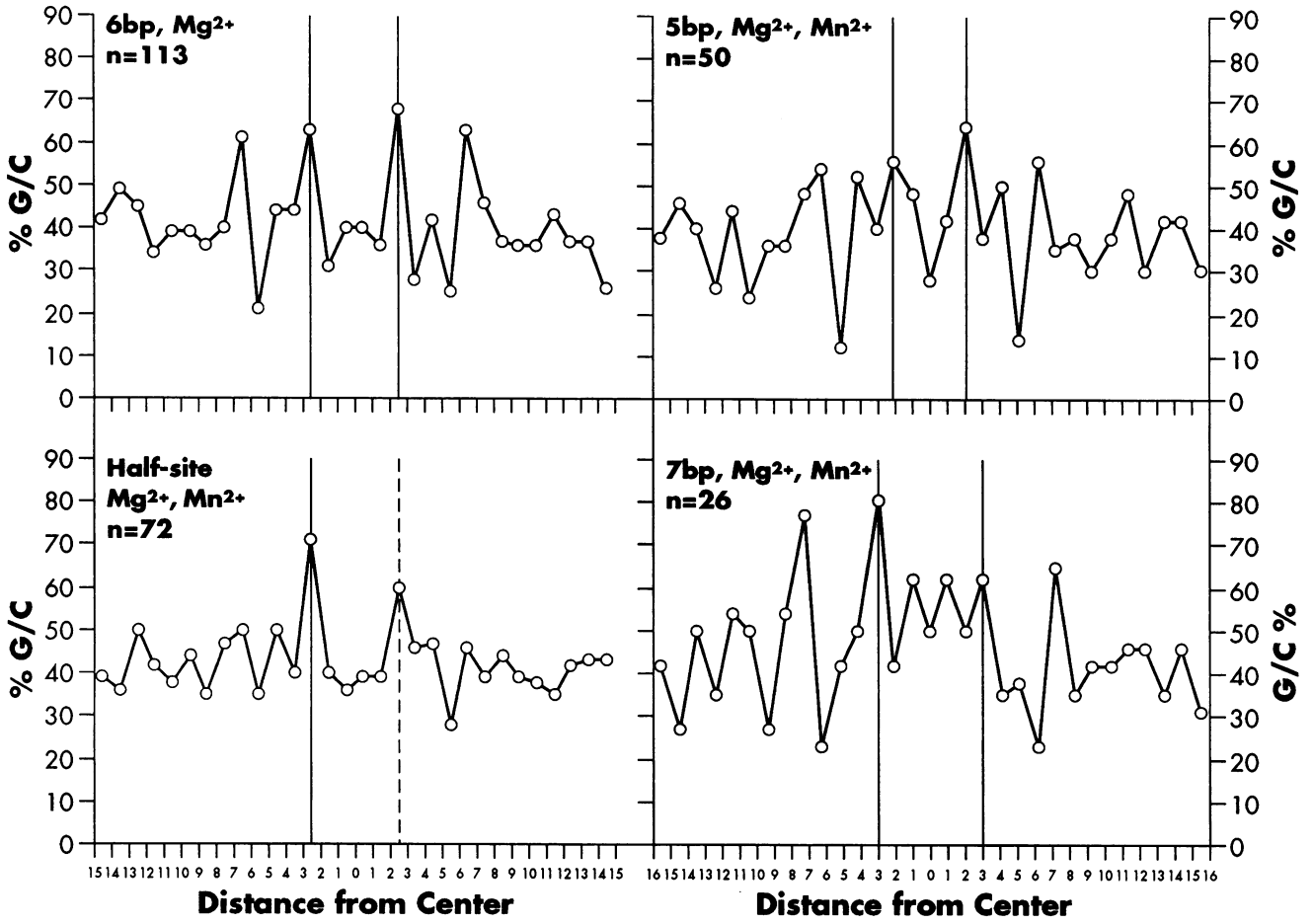

FIG. 2. G/C inflection patterns. The integration sites were aligned relative to their duplication centers, and the percent G/C content for each position was calculated and graphed. For the 5-, 6-, and 7-bp patterns, the positions attached to the LTR termini are bisected by the vertical lines, with the pGEM-3 T7 ends on the left. (All four sets are a combination of R35, R33, and R55. For the 7-bp pattern, four sites from reference 17 were included.) Except for the half-site reaction, the positions at and between the lines were duplicated. Half-site indicates all deletions and duplications larger than 8 bp (n = 36). Because each half-site product has two independent strand transfer sites, 72 sequence alignments are formed. For the half-site panel, the line on the left marks the target site attachment for both the T7 and Sp6 ends of the half-site products. The dashed line on the right indicates where the other LTR end would have been ligated for a 6-bp duplication. All the 6-bp duplications were cloned from Mg^{2+} reactions. The other sets are a combination of Mg^{2+} and Mn^{2+} recombinants in ratios determined from Table 1.

and nonpreferred sites, as were events from individual clones. An *Eco*RV cloning bias was not indicated, since multiple copies of a clone could be sequenced from one phage population but the frequency of finding a clone from a lambda population did not correlate to a preferred location. Additionally, manipulations of the integration reaction (Mg^{2+} versus Mn^{2+}) generated new site types at unique locations, a result not expected in the case of a systematic cloning bias.

A third possible bias could arise during the repair of the gaps in the host DNA and the overhanging 5′ ends of the viral DNA that are produced by strand transfer of the LTR plasmids into the lambda DNA. The repair process is assumed to be catalyzed by enzymes of the bacterial host after packaging and infection of the DNA from the integration reaction. This possible bias will be addressed in Discussion.

G/C inflection patterns. To gain a better understanding of the role that the host sequences (at, and immediately 5′ of, the insertion sites on both strands) have in concerted integration events, we analyzed the integration sites in the following manner. The integration sites were segregated into sets according to duplication structure and aligned relative to the center of the duplication. The percent G/C content was calculated for each position (Fig. 2). There was an indication of nucleotide sequence bias relative to the points of strand transfer. For the set of 6-bp Mg^{2+} duplications, peaks of G/C content were evident at the sites of strand transfer and 4 bp 5′ to the sites of strand transfer. When the Mn^{2+} 6-bp duplications were aligned (n = 33) (data not shown), a similar G/C pattern was observed. Sets of 50 5-bp and 26 7-bp duplicated sites were generated by combining the Mg^{2+} and Mn^{2+} data (Fig. 2). Both the 7- and 5-bp sites had peaks and troughs in G/C content, relative to the sites of strand transfer, similar to those observed with the 6-bp site pattern.

The statistical significance of the G/C pattern associated with the 6-bp duplicated sites was previously shown with a set size of 68 (10). A similar analysis was done with the 5-bp Mg^{2+} and Mn^{2+} set. The chi-square test on the 5-bp set also showed that the G/C values differed significantly from a homogeneous binomial distribution around the expected 45.7% average G/C content of the lambda nonessential region (χ^2_{29} = 92.49; $P <$

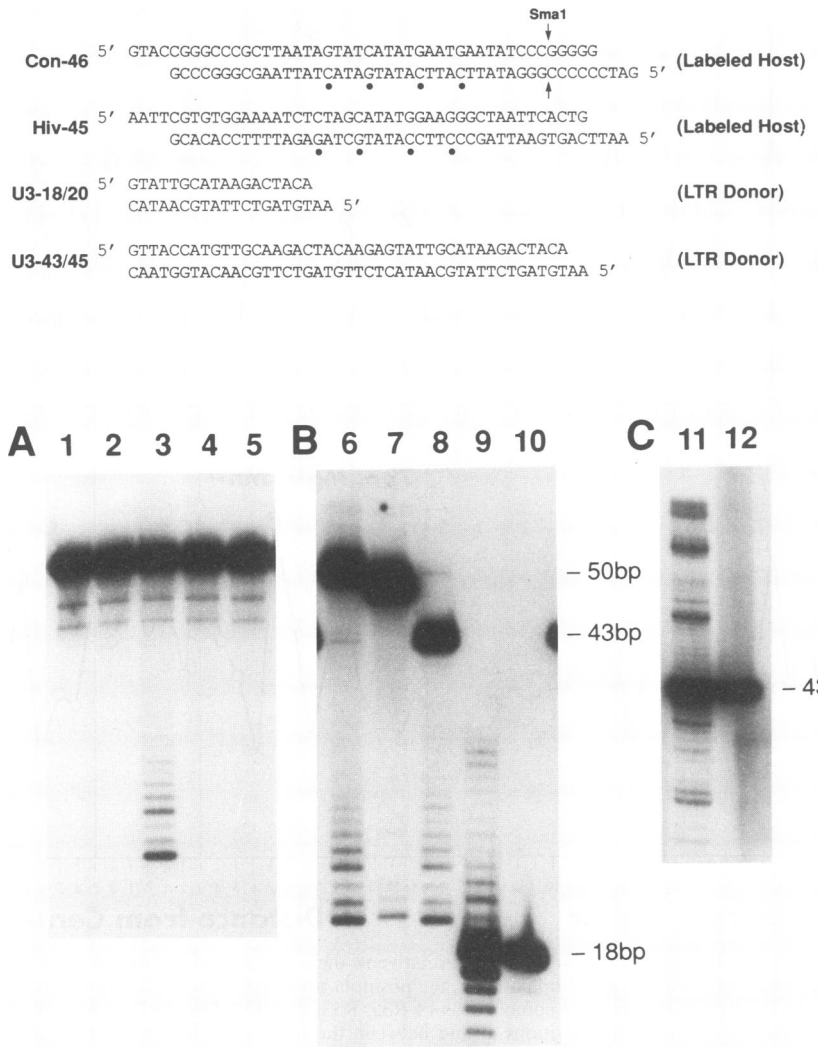

FIG. 3. (Top) Oligonucleotide substrates. Four substrates were synthesized (Operon, Inc.). The single-stranded oligonucleotides were purified by high-pressure liquid chromatography or on denaturing gels. For 3′ end labeling, the appropriate oligonucleotides were annealed and filled in with Klenow polymerase, labeling with [^{32}P]dCTP or [^{32}P]dTTP (Con-46 and Hiv-45, respectively). Con refers to the 6-bp G/C consensus pattern of Fig. 2. The dots below Con-46 and Hiv-45 denote G/C pairs which match the G/C peaks of the 6-bp pattern. The *Sma*I site on Con-46 is indicated. U3-18/20 and U3-43/45 are based on the avian U3 terminus and can serve as unlabeled LTR donors to labeled Con-46 and Hiv-45 or, when 5′ labeled on the recessed CA strand, as donors for integration into other labeled LTR molecules. (Bottom) Autoradiographs of representative experiments. (A) 3′-labeled Con-46 reacted with various unlabeled donors. Lanes: 1, Con-46 alone; 2, Con-46 with IN; 3, Con-46 with IN and unlabeled U3-18/20; 4, Con-46 with IN and single-stranded U3-18; 5, Con-46 with IN and a double-stranded HIV U5 donor, 19/21. (B) Sizing reaction products. Lanes: 6, 3′-labeled Con-46, IN, and unlabeled U3-18/20; 7, 3′-labeled Hiv-45, IN, and unlabeled U3-18/20; 8, Con-46 cut with *Sma*I, IN, and unlabeled U3-18/20; 9, 5′-labeled U3-18/20 and IN; 10, 5′-labeled U3-18/20. (C) 5′-labeled U3-43/45. Lanes: 11, U3-43/45 with IN; 12, U3-43/45 alone.

0.001). A set of 50 randomly picked sites from the nonessential region did not differ significantly ($\chi^2_{29} = 28.57$; $P = 0.49$). The G/C pattern results combined with the fact that preferred insertion sites biased toward specific duplication sizes were found indicates a correlation between the duplication size and the sequence of the host DNA for the full-site reaction.

Additional recombinants were cloned from the lambda system which appeared to be the products of uncoupled attack by the LTR termini (half-site reaction). The half-site reaction set represented a mixed population of integration events, which when cloned by the *Eco*RV method indicated that the termini of the integration products had not been closely juxtaposed during strand transfer. The set included duplications larger than 8 bp and all deletions. Overall these products were cloned at a rate of 13%, with a total of 36 events constituting the set (Table 1). Interestingly, with Mn^{2+}, duplications and deletions were found to range below 50 bp in size, while larger uncoupled products were found with Mg^{2+} (Table 1). By aligning the target DNA at the T7 and SP6 ends of the recombinants (10) and then combining these alignments such that the sites of transesterification overlapped, the half-site set ($n = 72$) was formed (Fig. 2). The pattern had a strong G/C peak at the site of strand transfer and some features of the 6-bp pattern, but a high G/C content 4 bp 5′ to the site of strand transfer was not present (Fig. 2). That the G/C pattern of the apparent half-site reactions was not a symmetrical half of that

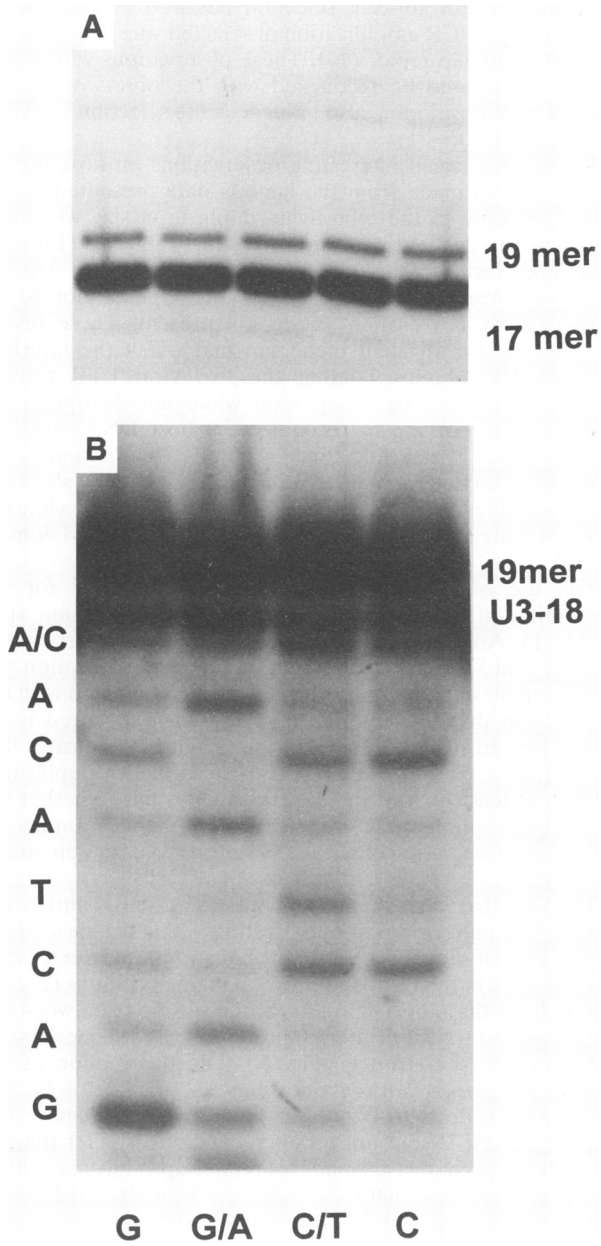

FIG. 4. (A) Scaled-up reaction for the isolation of the 19-mer product. Results for five 40-μl reaction mixtures with Mg^{2+} and the standard oligonucleotide reaction conditions are shown. The reactions were run on a 12% denaturing gel (sample of lanes shown, 4-h exposure), and the 19-mer was isolated. Also indicated is the position of the 17-mer. (B) After isolation, the 19-mer was subjected to Maxam and Gilbert chemical degradation (G, G/A, C/T, and C reactions were performed). The 19-mer and contaminating unreacted U3-18 are indicated at right. Just below U3-18 is the first degradation product, which has the possibility of being either an A or a C nucleotide, coming for the host molecule. Below this band, the LTR sequence of the donor molecule is apparent. The exposure time was 3 weeks.

of the full-site reactions may indicate a difference in host sequence recognition for the half-site and full-site reactions.

Site selection in the context of target oligonucleotide assays. We wanted to determine if AMV IN could recognize a consensus G/C pattern internalized in a larger double-stranded target oligonucleotide (46 bp) (Con-46) (Fig. 3, top) with a short double-stranded LTR oligonucleotide (18 and 20 bp; designated U3-18/20) as donor. A reaction with a 3′-end-labeled non-LTR target molecule and an unlabeled LTR donor allows for analysis of target site sequence usage within the target molecule (16, 20). The size of the reaction products will be the length of the LTR donor plus the number of sequences picked up from the target. Also, if the donor LTR oligonucleotide is 5′ end labeled and the only molecule in the reaction, it will also serve as the target (6, 17).

When unlabeled U3-18/20 was incubated with labeled Con-46 (Fig. 3, lane 3), using AMV IN, products running significantly below unreacted Con-46 (50 bp with fill-in; lane 1) were observed. The reactions in lanes 1 through 8 of Fig. 3 were run with Mg^{2+}. A control of IN alone did not produce these products (Fig. 3, lane 2), nor did a control of IN with donors consisting of single-stranded U3-18 or a double-stranded human immunodeficiency virus (HIV) U5 LTR oligonucleotide (lanes 4 and 5, respectively). The reaction of unlabeled U3-18/20 with another G/C consensus target (Hiv-45) (Fig. 3, top) produced a banding pattern (lane 7) similar to that observed with the Con-46 reaction (lanes 3 and 6). Con-46 with one end removed by *Sma*I digestion (a 43-mer target) produced a pattern similar to that produced by full-length Con-46 and showed no significant products in the range of 45 to 47 bp (the band at 50 bp is contaminating full-length Con-46) (Fig. 3, lane 8). Lane 10 of Fig. 3 contains unreacted 5′-labeled U3-18/20, indicating that the fast-migrating products from targets Con-46 and Hiv-45 were in the 20-bp range. No significant quantities of strand transfer products in the 45- to 47-bp range were observed (Fig. 3). Additionally, there was not significant integration into the 5′ ends of the molecules (60-bp range). These results indicated that the 6-bp G/C patterns did not serve as preferred host sites in the context of the oligonucleotide assays. Instead, 3′ ends were preferred in a manner similar to the usage of the LTR ends during the cleavage reaction. Whether the preference was related to LTR inverted repeat sequence recognition was next addressed.

LTR inverted repeat sequences not preferred as a target. When 5′-end-labeled U3-18/20 was reacted with IN, the most intense product was an apparent 19-mer (Fig. 3, lane 9) (Mn^{2+} was the cofactor; a shorter exposure showed a cleanly running band at that position) (also see Fig. 4A). The simplest explanation for the generation of a 19-mer reaction product was that it occurred by strand transfer into another U3-18/20 molecule, one nucleotide from either 3′ end (Fig. 3). Selection of the recessed catalytic LTR terminus for strand transfer would ligate an A nucleotide onto the donor U3-18 strand (Fig. 3, top). Selection of the blunt-ended noncatalytic terminus would ligate a C nucleotide onto U3-18. A scaled-up reaction was run with Mg^{2+} as the cofactor (Fig. 4A), and the 19-mer was isolated and sequenced by chemical degradation (Fig. 4B). Sequencing results indicated that both ends were used, with the possibility of the noncatalytic terminus being preferred. The extent of the C/T or C reaction appeared to be greater than that of the G/A reaction. In our interpretation of the sequence data, we assumed that there was nonspecific background degradation in the G lane.

Preference for the noncatalytic end was also indicated by the reaction product running just below unreacted U3-18 (Fig. 4A). This 17-bp product would be produced by strand transfer into the recessed LTR 3′-OH end ("labeled kickout product") (2). However, since the 17-mer was of considerably lower intensity than the 19-mer, again selection of the noncatalytic terminus was indicated (Fig. 4A). We confirmed that the strand transfer bias for the noncatalytic terminus of the

U3-18/20 substrate was due to sequence and not to end structure (blunt versus recessed) by synthesizing a substrate with identical 5′ overhangs on both the catalytic and noncatalytic ends. Again, with the new substrate, the 3′-OH end of the noncatalytic terminus was preferred, indicating that sequence and not end structure was likely responsible for the preference for the noncatalytic LTR terminus (data not shown).

Results obtained by using the larger LTR donor substrate U3-43/45 also indicated that the catalytic LTR terminus of this molecule did not serve as a preferred strand transfer site. With Mn^{2+} as the cofactor, the reaction pattern of the 5′-end-labeled U3-43/45 substrate serving as both donor and target (Fig. 3, lanes 11 and 12) was considerably different from that observed with U3-18/20 (Fig. 3, lanes 9 and 10). For this 5′-end-labeled U3-43/45 substrate, the 3′ ends did not appear to be preferred (no strong bands ran just above unreacted 43-mer). Instead, several strong internal and 5′ regions were preferred as strand transfer sites. This result again suggests that the terminal catalytic LTR sequences of U3-43/45 did not serve as a preferred site of integration.

Disintegration activity associated with IN (25) could provide an explanation for the preference for 3′ ends as sites of strand transfer for the U3-18/20, Con-46, and Hiv-45 substrates. Strand transfer into the central region of a duplex oligonucleotide would produce a stable Y-shaped molecule which could serve as a substrate for disintegration. Strand transfer into the 3′ ends of a target oligonucleotide would produce unstable molecules that would not serve as disintegration substrates. These unstable products could accumulate during the reaction and would thus appear as preferred sites for the forward integration reaction. This does not seem to be the case, since for U3-43/45, internal regions and regions nearer the 5′ end of this molecule were preferred sites.

DISCUSSION

We have previously shown that concerted integration of a virus-like DNA donor by AMV IN into naked lambda DNA is nonrandom. In this study further evidence for nonrandom selection on naked DNA was indicated by the cloning of preferred insertion sites that had sustained multiple insertions down to the nucleotide. The 6-bp event indicated a 10^9 bias for this host site over other possible sites.

That the observed biases were not due to contamination or amplification during cloning of the integration sites was addressed in Results. Repair of the gaps in the integration products by bacterial host enzymes also does not appear to be responsible for the nonrandom aspects of our integration system. A previous comparison of our results with those for MLV integration, also into naked lambda DNA (1), indicated that site selection for these two retrovirus systems was not identical (10). Since bacterial host enzymes were used for gap repair in both systems but the distribution of selected sites was different, control of site selection does not appear to be strongly influenced by gap repair.

The influence of sequences flanking the duplicated sites also appears to be different for AMV (10) and MLV (22) integration with naked DNA as the target. The G/C content graphs of Fig. 2 consistently show a strong bias for A/T base pairs at positions 3 bp outside the duplicated sequences. For the MLV sites, a strong preference for A/T base pairs was observed 2 bases out from each side of the 4-bp MLV duplications (22). These different biases were observed with systems that again relied on bacterial host enzymes to repair the gaps. The results again indicate that control of site selection is at the level of integration. The observation that AMV IN is the major controlling factor for site selection has also been made by others using a PCR amplification of selected sites that does not depend upon gap repair (19). These observations with naked DNA will have to be reconciled with the observations that nucleosomal positioning also influences site selection for MLV integration (22).

Several additional interesting observations on host site selection can be made from the lambda data presented in this report. Fidelity to the 6-bp avian duplication size was influenced by the divalent metal ion present (Table 1). However, preference for the symmetrical G/C content of the 5′ flanking host sequences at the duplicated site appears not to be influenced by the divalent metal ion present. These results suggest that the divalent metal ion may affect the interface between IN subunits holding the two viral LTR termini together for concerted integration while not disrupting the ability of the enzyme to recognize the host DNA. Such an interpretation is plausible, since a putative metal-binding region (18) has been mapped to a central region of IN shown to be responsible for both catalysis and multimerization, while the DNA binding ability of the protein maps to a carboxyl domain (7, 12).

The use of Mn^{2+} as a metal cofactor allowed for the generation of larger numbers of aberrant duplication sizes (Table 1). Analysis of the G/C patterns of the aberrant site types (7 and 5 bp [Fig. 2]) indicated that each duplication size had a unique sequence pattern. This result, combined with the observation that the multiple insertion sites appeared to be biased toward specific duplication sizes (Fig. 1), indicates that the sequence of the host site may also influence duplication fidelity, at least for this purified in vitro system. Whether the host sequence influences site selection and duplication fidelity for cellular integration of viral DNA into chromatin remains to be tested.

The fact that each duplication size (5, 6, and 7 bp) had a unique G/C pattern was also interesting in that it indicated the importance of sequences at the sites of strand transfer relative to sequences in the third and fourth flanking positions (Fig. 2). For at least the 6-bp G/C pattern, the positions between the sites of strand transfer (internal host duplicated sequences) do not appear to be strongly selected. The distribution of the lambda site types, with duplications as small as 4 bp but deletions only down to 12 bp, is noteworthy in this regard. The fact that AMV IN did not produce deletions smaller than 12 bp, while duplications as small as 4 bp were produced, could be interpreted to mean that there is less steric hindrance 3′ to the site of strand transfer relative to sequences 5′ during host site binding by IN. We find these trends in the lambda data significant in that they indicate a similarity between recognition of the host site DNA and recognition of the LTR 3′-OH ends. For the LTR catalytic termini, a small amount of sequence immediately 5′ to the conserved CA moiety defines recognition specificity for trimming and strand transfer (2, 5). For the HIV trimming reaction, adduct interference experiments indicate that HIV IN does not make tight contacts with the terminal sequences 3′ to the conserved CA moiety (3).

We directly tested whether 6-bp G/C nucleotide symmetry could be preferentially recognized by AMV IN within the context of a target oligonucleotide (16, 20) with an LTR oligonucleotide as the donor. The data showed that IN did not recognize the internalized symmetrical target site on the oligonucleotide but rather that the 3′ ends served as preferred sites for strand transfer. Others have noted that HIV IN can catalyze strand transfer within 2 bp from the 3′ ends on a duplex oligonucleotide (16), but a preference for these sites was not observed. The AMV IN preference for 3′ ends as sites

of strand transfer on a duplex oligonucleotide is significant because again it indicated a similarity between host site recognition and LTR recognition. Testing whether the preference for 3′ ends was related to recognition of LTR sequences showed that this was not the case (Fig. 3 and 4). Thus, while it may be possible that one catalytic site is responsible for trimming and strand transfer (8, 21), recognition of the host site DNA and recognition of the LTR termini do not appear to be identical in terms of substrate sequence effects. This conclusion is also supported by the fact that G/C sequence bias patterns for concerted integration (Fig. 2) do not resemble the sequences of the avian LTR termini.

Why the 6-bp G/C pattern was not a preferred site in the context of the oligonucleotide assays was not clear. The lack of preference may be related to there being different sequence effects for the half-site and full-site integration reactions. This possibility was indicated by the G/C pattern of the apparent half-site reaction in the lambda system (Fig. 2) not being a symmetrical half of the full-site patterns. Although there still appeared to be prominent selection for G/C base pairs at the site of strand transfer, selection for G/C base pairs at the position 4 bp 5′ to the site of strand transfer was lost. In another assay measuring the half-site integration of the R35 LTR plasmid into other R35 plasmids, we found a G/C pattern that was more a symmetrical half of the full-site pattern (15). Again, however, there did not appear to be selection for G/C base pairs at the fourth position 5′ to the site of strand transfer. The oligonucleotide assays are likely dominated by the half-site reaction (21). Whether the lack of preference for the 6-bp G/C pattern in the oligonucleotide assays was related to these observations or to other possibilities, such as oligomerization states of AMV IN on the substrates and the presence of ends on the host target substrates, remains to be investigated.

ACKNOWLEDGMENTS

Technical assistance for protein purification and sequencing was provided by A. C. Vora, W. Zeh, and M. McCord. Statistical analysis was performed by George Vogler, Division of Biostatistics, Washington University Medical Center.

This work was supported by grants from the National Institutes of Health.

REFERENCES

1. **Brown, P., B. Bowerman, H. Varmus, and M. Bishop.** 1987. Correct integration of retroviral DNA in vitro. Cell **49**:347–356.
2. **Bushman, F., and R. Cragie.** 1991. Activities of human immunodeficiency virus (HIV) integration protein *in vitro*: specific cleavage and integration of HIV DNA. Proc. Natl. Acad. Sci. USA **88**:1339–1343.
3. **Bushman, F., and R. Cragie.** 1992. Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites. Proc. Natl. Acad. Sci. USA **89**:3458–3462.
4. **Bushman, F., T. Fujiwara, and R. Cragie.** 1990. Retroviral DNA integration directed by HIV integration protein *in vitro*. Science **249**:1555–1558.
5. **Colicelli, J., and S. Goff.** 1988. Sequence and spacing requirements of a retrovirus integration site. Mol. Biol. **199**:47–59.
6. **Craigie, R., T. Fujiwara, and F. Bushman.** 1990. The IN protein of moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro*. Cell **62**:829–837.
7. **Engleman, A., F. Bushman, and R. Craigie.** 1993. Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. EMBO J. **12**:3269–3275.
8. **Engleman, A., K. Mizuuchi, and R. Craigie.** 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell **67**:1211–1221.
9. **Fitzgerald, M., A. Vora, and D. Grandgenett.** 1991. Development of an acid-soluble assay for measuring retrovirus integrase 3′-OH terminal nuclease activity. Anal. Biochem. **196**:19–23.
10. **Fitzgerald, M., A. Vora, B. Zeh, and D. Grandgenett.** 1992. Concerted integration of viral DNA termini by purified avian myeloblastosis virus integrase. J. Virol. **66**:6257–6263.
11. **Fujiwara, T., and R. Craigie.** 1989. Integration of mini-retroviral DNA: a cell-free reaction for biochemical analysis of retroviral integration. Proc. Natl. Acad. Sci. USA **86**:3065–3096.
12. **Gent, D., C. Vink, A. Groeneger, and R. Plasterk.** 1993. Complementation between HIV integrase proteins mutated in different domains. EMBO J. **12**:3261–3267.
13. **Gent, D., C. Vink, A. Groeneger, and R. Plasterk.** 1993. Identification of amino acids in HIV-2 integrase involved in site-specific hydrolysis and alcoholysis of viral DNA termini. Nucleic Acids Res. **21**:3373–3377.
14. **Goff, S.** 1992. Genetics of retroviral integration. Annu. Rev. Genet. **26**:525–542.
15. **Grandgenett, D., R. Inman, A. Vora, and M. Fitzgerald.** 1993. Comparison of DNA binding and integration half-site selection by avian myeloblastosis virus integrase. J. Virol. **67**:2628–2636.
16. **Hong, T., E. Murphy, J. Groarke, and K. Drlica.** 1993. Human immunodeficiency virus type 1 DNA integration: fine-structure target analysis using synthetic oligonucleotides. J. Virol. **67**:1127–1131.
17. **Katz, R., G. Merkel, J. Kulkosky, J. Leis, and A. Skalka.** 1990. The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. Cell **63**:87–95.
18. **Khan, E., J. Mack, R. Katz, J. Kulkosky, and A. Skalka.** 1990. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res. **19**:851–860.
19. **Kitamura, Y., Y. M. H. Lee, and J. Coffin.** 1992. Nonrandom integration of retroviral DNA *in vitro*: effect of CpG methylation. Proc. Natl. Acad. Sci. USA **89**:5532–5536.
20. **Leavitt, A. D., R. B. Rose, and H. E. Varmus.** 1992. Both substrate and target oligonucleotide sequences affect in vitro integration mediated by human immunodeficiency virus type 1 integrase protein produced in *Saccharomyces cerevisiae*. J. Virol. **66**:2359–2368.
21. **Mizuuchi, K.** 1992. Polynucleotidyl transfer reactions in transpositional DNA recombination. J. Biol. Chem. **267**:21273–21276.
22. **Pryciak, P. M., A. Sil, and H. E. Varmus.** 1992. Retroviral integration into minichromosomes *in vitro*. EMBO J. **11**:291–303.
23. **Sandmeyer, S. B., L. J. Hansen, and D. L. Chalker.** 1990. Integration specificity of retrotransposons and retroviruses. Annu. Rev. Genet. **24**:491–518.
24. **Varmus, H.** 1983. Retroviruses, p. 411–503. *In* J. Shapiro (ed.), Mobile genetic elements. Academic Press, New York.
25. **Vincent, K. A., V. Ellison, S. A. Chow, and P. O. Brown.** 1993. Characterization of human immunodeficiency virus type 1 integrase expressed in *Escherichia coli* and analysis of variants with amino-terminal mutations. J. Virol. **67**:425–437.
26. **Vora, A., M. Fitzgerald, and D. Grandgenett.** 1990. Removal of 3′-OH-terminal nucleotides from blunt-ended long terminal repeat termini by the avian retrovirus integration protein. J. Virol. **64**:5656–5659.
27. **Whitcomb, J., and S. Hughes.** 1992. Retroviral reverse transcription and integration. Annu. Rev. Cell Biol. **8**:297–304.