Concerted Integration of Viral DNA Termini by Purified Avian Myeloblastosis Virus Integrase

MICHAEL L. FITZGERALD, AJAYKUMAR C. VORA, WILLIAM G. ZEH,
AND DUANE P. GRANDGENETT*

*Institute for Molecular Virology, St. Louis University Medical Center,
3681 Park Avenue, St. Louis, Missouri 63110*

Received 7 May 1992/Accepted 7 August 1992

Concerted integration of retroviral DNA termini, which produces a characteristic duplication of sequences at the integration site and formation of the proviral state, is a necessary step of the retroviral life cycle. We investigated the pairwise integration reaction catalyzed by purified avian retrovirus integrase by measuring the response to solution parameters and how the sequences of the viral termini, which comprise the avian imperfect inverted repeat, affect the reaction. When we optimized the reaction, an efficiency was achieved which approached that measured in systems using cytoplasmic extracts from virus-infected cells. The response of purified avian integrase to solution parameters was similar to that of the integration activity derived from cellular extracts. For strand transfer, the U3 viral terminal sequences were preferred to those of the U5 termini, a result we previously showed for the trimming reaction. That the sequence preference was the same for trimming and strand transfer may be further evidence that only one catalytic site is used for both reactions. A significant number of integration sites were sequenced. Interesting trends were found for the fidelity of the host duplications to the avian 6-bp duplication size, the clustering of the integration sites in the nonessential region of the lambda host DNA, and the sequence characteristics of the duplication sites.

As part of the retroviral life cycle, there is a transition from the free infectious RNA particle to the integrated provirus, which is essentially a genetic element within the host cell genome. Integration appears to be necessary for efficient virus propagation in cell culture, and through mutagenesis the virally encoded integrase (IN) has been shown to be essential for this process (8, 22, 24, 27). In vitro, purified IN has been shown to be capable of catalyzing the major steps of the integration reaction (7, 17). Because integration appears to be a necessary step in the life cycle of retroviruses and there is likely no essential homologous cellular reaction, inhibitors of IN hold promise of therapeutic potential for retroviral infection.

The current integration model has the reaction occurring by a two-step mechanism. First, the blunt ends of the linear viral DNA have two nucleotides removed from the 3'-OH ends in an endolytic reaction. This reaction exposes the highly conserved 5' CA-OH 3' moieties which will delineate the viral ends in the provirus. After 3'-end processing, these 3'-OH ends are joined to the 5' ends of the host site. This second step, designated strand transfer, produces the gapped proviral intermediate which has been assumed to be repaired by cellular enzymes (2, 14), but recent evidence indicates a possible role for IN (5). During infection, there appears to be a spatial and temporal separation of the two steps. The 3'-end processing is nearly complete during the transition of the viral DNA through the cytoplasm, while strand transfer must occur in the nucleus (2, 26). In vitro, 3'-end processing and strand transfer have also been decoupled (13).

A variety of in vitro assay systems have been developed to study the two reactions (1a, 4, 7, 10, 13, 18, 31). To study 3'-end processing, we have employed a long terminal repeat (LTR) plasmid-based system and avian myeloblastosis virus

(AMV) IN purified from virions (12). It was shown that purified AMV IN is capable of releasing 50% or more of the dinucleotide from input LTR plasmid DNA under optimal conditions. Additionally, preferential 3'-end processing of avian retroviral U3 LTR sequences over those of the U5 LTR was demonstrated (12, 33).

In this study, we used the plasmid-based LTR substrates in a genetic assay (1a) to measure the strand transfer activity of purified AMV IN. We show here, as have others (17), that for reaction conditions which are optimal for integration of a processed plasmid, a plasmid with blunt LTR ends is not effectively integrated. However, it is shown that as for the 3'-end processing reaction, the preference of AMV IN for U3 substrates also holds for strand transfer. Additionally, it was observed that Mg^{2+} induced a higher level of concerted (pairwise) LTR terminus integration than Mn^{2+} . This is of interest in that assays employing LTR duplex oligonucleotide substrates have indicated that Mn^{2+} induces a higher level of strand transfer activity (7, 17).

During this study, we sequenced a significant number of recombinants produced by the various viruslike plasmids and thus were able to analyze the sequence characteristics of a more substantial set of host sites. We found several interesting trends and compared our results with those found for in vitro integration of Moloney murine leukemia virus (MoMLV) DNA endogenous to the subcellular 160S nucleoprotein complex into naked host DNA (1a, 23).

MATERIALS AND METHODS

To detect integration, an assay was developed in which *Nde*I-linearized LTR plasmids carrying the *supF* gene are integrated into lambda gtWES (1a, 13, 31). Lambda gtWES has amber mutations in essential genes and can grow in *SupF*⁻ bacteria only if an LTR plasmid has integrated into the nonessential region of the genome. Comparison with the amount of phage able to grow on a *SupF*⁺ strain allows

* Corresponding author.

calculation of the total integration events. To determine the structures of LTR-lambda joints, the integrated plasmids are subcloned out of lambda and sequenced as double-stranded plasmids.

Assay conditions. Reactions were run in 30- μ l volumes with 0.5 μ g of *Nde*I-linearized LTR plasmid and 0.5 μ g of nonconcatemered lambda DNA. For optimal efficiency, final solution conditions included 100 mM NaCl, 20 mM Tris-hydrochloride (pH 7.5), 10 mM dithiothreitol, 5 mM $MgCl_2$, 10% dimethyl sulfoxide (DMSO), 5% polyethylene glycol, 1% glycerol, and 0.05% Nonidet P-40. These solution conditions were used for all of the experiments presented, except as noted otherwise in the figure legends. The DNAs, in reaction buffer, were mixed with purified AMV IN on ice and incubated for 10 min; this was followed by addition of polyethylene glycol and incubation at 37°C for 30 min. Reactions were stopped by addition of 120 μ l of 6 mM EDTA, 7.5 μ l of 10% sodium dodecyl sulfate, and 1 μ l of a 10-mg/ml solution of proteinase K and incubation for 1 h at 37°C. To the reactions, 20 μ l of 3 M sodium acetate (pH 7.0) and 15 μ g of *Saccharomyces cerevisiae* tRNA were added; this was followed by phenol-chloroform-isoamyl alcohol (24:24:1) extraction and two ether extractions. The DNA was ethanol precipitated, washed twice with 70% ethanol, suspended overnight (4°C) in 5 μ l of 10 mM Tris-hydrochloride–0.1 mM EDTA (pH 7.5), and packaged by the Gigapack Plus II System (Stratagene). All packaging of lambda and subsequent plating on bacterial cells were performed under nonlimiting plating conditions.

LTR plasmid construction. A 65-bp double-stranded LTR synthetic insert was cloned into the *Eco*RI site of pGEM-3 (12). The *supF* gene (469 bp) from pAN7 (gift from J. Coffin) was cloned into the above plasmid at the *Hind*III site. The LTR insert has an *Nde*I site between the U3 and U5 terminal sequences, and the resulting plasmid is designated R35 (see Fig. 1). Digestion of R35 with *Nde*I produces a linear molecule which mimics the retroviral genome subsequent to the trimming reaction (13). During the cloning of R35, a plasmid was isolated with the LTR insert in the opposite orientation. It was thus possible to clone R33 and R55 plasmids with only U3 or U5 sequences, respectively. We previously constructed pAV-DraI, a pBR322-based plasmid which has a *Dra*I site instead of an *Nde*I site at the junction between the LTR termini; linearization with *Dra*I produces a molecule which mimics the blunt-ended viral genome, with the exception of an additional A-T base pair at each terminus (33).

Plasmids were purified by using Qiagen pack 500 columns and further purified on a 5 to 20% sucrose gradient to isolate the supercoiled form and remove any contaminating small nucleic acids. Digestion with *Nde*I (Bethesda Research Laboratories [BRL]) was done for 1 h at 2 U/ μ g of purified DNA. It is important that the *Nde*I preparations contain no contaminating exonucleases, because even partial degradation of the LTR 3' ends severely reduces the activity of the cut plasmid for integration (data not shown). *Nde*I appeared to be sensitive to the large inverted repeat of R33 and R55 created during cloning; the enzyme inefficiently cut the supercoiled form, and relaxation with topoisomerase 1 (BRL) was necessary before effective cutting with *Nde*I.

Lambda DNA preparation. Lambda gtWES·lambda B was purchased initially from BRL. Lambda gtWES was propagated in LE-392 cells (Stratagene), and virions and DNA were purified as previously described (28). The resulting DNA was further purified over a 10 to 40% sucrose gradient before being used in integration reactions. Addition-

ally, each preparation was screened for lack of growth on *SupF*⁻ cells.

Integrase purification. IN was purified nearly to homogeneity from AMV (15, 19). The molecular weight of native dimeric IN is 64,000 (15).

Scoring of integration. Bacterial strains LE-392 (*SupF*⁺) (Stratagene), LG-75 (*SupF*⁻), and MC-1061 (*SupF*⁻) were used for growth of lambda phage. The latter two strains were obtained from H. Huang (Washington University). The total number of integration events was calculated by the formula of Brown et al. (1a), except for factor E, which was omitted (factor E corrected for insertion of the 10-kb viral DNA into lambda gtWES, whose packaging is sensitive to insert size). Our inserts are 3.4 kb long and should have a minimal positive effect on packaging. We routinely used MC-1061 (*SupF*⁻) cells to determine the number of plasmids (R35, etc.) integrated into lambda gtWES. MC-1061 consistently produced two to three times more plaques than did LG-75.

Host site analysis. To determine host site structure, individual plaques were picked and amplified in the MC-1061 strain. Lambda DNA was purified (Qiagen), digested with *Eco*RV (Promega), and ligated (T4 ligase [BRL]) under dilute conditions. HB101 cells (BRL) were transformed, and plasmids from ampicillin-resistant colonies were sequenced by using the dideoxy method and primers that anneal in the polylinker region of pGEM-3.

The Sequence Analysis Software Package (Genetics Computer Group, University of Wisconsin) was used to locate the integration sites in the lambda genome and calculate the A-T composition of the nonessential region of the genome. The host sequences of the integration sites were aligned relative to the center of the duplicated bases, and the percent G-C composition was calculated for each position. Additional sites were selected from the lambda genome as described in the legend to Fig. 5 and were aligned in a similar manner for comparison with the integration set. The statistical relevance of the variation in A-T composition was assessed by a chi-square test. Since there were two outcomes (A-T or G-C) and 30 positions, the test had 29 degrees of freedom. A *t* test was used to determine whether the average composition of the alignments differed significantly from the expected average of the major nonessential region of lambda gtWES·lambda B (45.7% G+C). A list of the sequenced host sites and selected sites is available upon request.

RESULTS

Strand transfer reaction optimization. We have developed an integration assay that incorporates elements of systems previously used in other laboratories (1a, 13, 31). AMV IN purified from virions provides the integration activity, and *Nde*I-linearized LTR plasmids are used as viral DNA substrates. Initial reaction conditions used for integration of the wild-type R35 plasmid (Fig. 1) were from the study of Fujiwara and Craigie (13). We found that cell extracts from uninfected chicken embryo fibroblasts did not reliably stimulate integration (data not shown). Removal of RNase A and glycerol (beyond that brought in by IN) did not affect efficiency, but removal of Nonidet P-40 and DMSO (minimal reaction buffer) resulted in a 90% decrease (compare mix reactions of Table 1). DMSO has previously been shown to affect the MoMLV integration reaction (13). The preincubation on ice could be reduced to 10 min without a negative effect. The reaction was not sensitive to salt between 25 and 200 mM (Fig. 2B), but activity dropped off sharply above 250

FIG. 1. LTR plasmid constructs. Three plasmids (R35, R33, and R55) were constructed by initially inserting a 65-bp synthetic avian retrovirus LTR oligomer into the *EcoRI* site and an *Escherichia coli* *supF* gene into the *HindIII* site of pGEM-3. R35 has the terminal 24 bp from the U3 LTR and 36 bp from the U5 LTR flanking an *NdeI* site. When cut with *NdeI*, the plasmid is linearized and possesses recessed 3' termini which mimic the trimmed retroviral termini. R33, with two identical 24-bp U3 termini, and R55, with two identical 36-bp U5 termini, were cloned by using R35 and another pGEM-3 plasmid in which the 65-bp LTR insert was in the opposite orientation to R35. Below the plasmid diagram are the sequences that comprise the imperfect inverted repeat, points of asymmetry are indicated by asterisks for R35. A third plasmid, pAV-DraI, has a 330-bp LTR insert with a *DraI* site at the center of the juxtaposed LTR termini. This plasmid also carries the *supF* gene. The sites where cleavage must occur on pAV-DraI before proper strand transfer can occur are indicated by arrows.

mM (data not shown). When the temperature was varied, an optimum of around 40°C was observed and the type of buffer had a minimal effect (Fig. 2D). Varying the pH of the Tris-hydrochloride or HEPES (N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid) buffer showed an optimum between 7.5 and 8.0, with a 50% decrease at pH 6.5 or 8.5 (data not shown). Varying the molar ratio of IN to R35 produced a linear relationship up to an 80:1 ratio of IN dimers to R35 molecules (Fig. 2A); at higher ratios, inhibition was observed (data not shown). The integration reactions plateaued after 60 min (Fig. 2C).

Reaction reproducibility. Three identical independent integration reactions were set up (with R35 as the substrate), and their integration efficiencies were compared to measure reproducibility. By infecting MC-1061 cells at nonsaturating levels of bacteriophage and plating them in duplicate, we were able to show an average of 1.4×10^5 integrations per 1×10^7 packagable phage genomes, with a variation of 3% (data not shown).

Specific activity. A value of 2.0×10^5 recombinants per 1×10^7 packagable phage genomes was found for R35 in the experiment presented in Table 1 (see Materials and Methods for the initial calculation). When the total amount of lambda DNA (0.5 μ g) was taken into account, a value of 2.3×10^8 recombinants was derived. This represents 0.17% of the

TABLE 1. Preference of U3 over U5 for strand transfer^a

Expt and DNA	Packaging efficiency (10 ⁷)	No. of integration events	Normalization to 10 ⁷ PFU
Expt 1			
R35 control	2.4	0	0
R35	4.2	8.3×10^5	2.0×10^5
R33	2.5	1.4×10^6	5.6×10^5
R55	3.0	4.2×10^5	1.4×10^5
Mixture	2.7	4.5×10^5	1.7×10^5
Expt 2, mixture	2.0	4.2×10^4	2.1×10^4

^a Integration efficiencies are presented for two experiments. Experiment 1 had an R35 control with no added IN, three identical reactions with either R35, R33, or R55 as the substrate (0.5 μ g), and a mixed reaction in which all three substrates were included (0.3 μ g each). The reaction conditions were similar to those described for Fig. 2, except that after the ice incubation, a 10-min incubation without Mg^{2+} at 25°C was included to allow IN time to equilibrate among the three substrates. Experiment 2 was another mixed reaction identical to experiment 1, except that no DMSO or Nonidet P-40 was added. The buffer was HEPES, and 64% of the 100 mM salt was KCl.

initial R35 substrate and 2.0% of the starting lambda DNA. A total of 722 ng of IN was included in this reaction, giving a molar ratio of 3.0×10^4 IN dimers per integration event.

At 1 mM Mn^{2+} , the enzyme appeared to be slightly more active than at 1 mM Mg^{2+} (Fig. 3A). At 5 mM, however, Mg^{2+} is strongly favored under optimal solution conditions. A concentration of 10 or 20 mM Mg^{2+} provided no further stimulation (data not shown). In a minimal buffer (no DMSO or Nonidet P-40, as in experiment 2 in Table 1), Mn^{2+} had higher activity at 5 mM but was still only 50% of the Mg^{2+} -dependent activity (data not shown). There was no significant loss of packagable lambda DNA at higher concentrations of Mn^{2+} ; thus, the apparent loss of activity could not be explained by activation of the nonspecific endonucle-

FIG. 2. LTR pairwise strand transfer optimization. Except for the parameters being varied, optimal assay conditions were used, with a 40:1 ratio of IN dimers to LTR molecules. For panel D, conditions were similar except that the pH of the respective buffers was 8.0 and the NaCl concentration for the HEPES curve was 133 mM and that for the Tris-hydrochloride curve was 160 mM. The curves in all four panels are representative examples of several experiments. Values given for integration events are 10^7 PFU on LE-392 (SupF⁺) cells.

FIG. 3. Divalent metal ion activity. (A) The efficiency of integration for R35 was compared for Mg²⁺ and Mn²⁺ at 1 and 5 mM concentrations. Optimal assay conditions were employed as described for Fig. 2A to C. Values given for integration events are per 10⁷ PFU on LE-392 cells. (B) Comparison of trimmed pFT-NdeI R35 with blunt-ended pAV-DraI as strand transfer substrates in the presence of either 5 mM Mg²⁺ or 1 mM Mn²⁺. Reaction conditions were similar to those described for panel A.

ase possessed by AMV IN rendering the lambda DNA unpackagable. The endonuclease, however, may nick at or near the CA-OH 3' moieties of the LTR termini, which are critical for substrate activity (18, 33).

For LTR pairwise integration, specific activity has the additional component of host site size duplication fidelity. In vivo, the size of host site duplication is correlated with the integrating virus and is nearly constant (29). For avian retroviruses, the duplication is 6 bp long. With purified AMV IN, however, in vitro integration does not have the same level of fidelity (17; see Table 3). In this study, the integration sites partitioned almost exclusively into a distribution of sites with 5-, 6-, and 7-bp duplication sizes. The 6-bp duplications predominated, and the aberrant 5-bp size was observed more frequently than the 7-bp size. It is of interest that the R55 plasmid may produce a greater number of 5-bp duplications relative to R35 and R33. We sequenced five recombinants produced in the presence of Mn²⁺ and found three 6-bp duplications, one 5-bp site, and interestingly, one 4-bp site. Additionally, apparent deletions of host site sequences were found (Table 2, experiment 1). Sequence analysis of the deletions indicated that approximately 500 bp

TABLE 2. Integration frequency and duplication distribution of R35, R33, and R55 in mixed reactions^a

Expt and LTR substrate	No. of duplications			No. of deletions	Total no. sequenced
	7 bp	6 bp	5 bp		
Expt 1					
R35	1	2	2	1	6
R33		6		2	8
R55		1			1
Expt 2					
R35	1	6	3		10
R33	1	6	1		8
R55		3			3

^a Individual plaques were picked from SupF⁻ plates of mixed reactions and amplified. Lambda DNA was purified, and the integrated plasmids were subcloned by the *EcoRV* method. Sequenced recombinants were categorized by plasmid species and size of host duplication. Apparent deletions of host site sequences were found in the reaction of experiment 1.

of the host sequence had been deleted for the R33 recombinants. For the R35 recombinant, host site structure was not clear and may have undergone recombinational events not related to integration. In all three recombinants, the viral LTR sequences were intact.

Comparison of LTR substrates. A series of LTR plasmid substrates was constructed which varied in the LTR sequences flanking the *NdeI* site (Fig. 1). As observed by others (17), a blunt-ended substrate, pAV-DraI, was used inefficiently (Fig. 3B). pAV-DraI has three nucleotides 3' of the CA sequence and must be trimmed before integration can occur (Fig. 1). It was integrated at 2 or 6% of the value of R35 in the presence of 5 mM Mg²⁺ or 1 mM Mn²⁺, respectively (Fig. 3B).

The terminal inverted repeat of avian retrovirus is imperfect (Fig. 1; asterisks denote points of asymmetry between the termini of R35), and we have previously observed that on isolated U3 or U5 large restriction fragments (181 and 155 bp, respectively) the U3 terminus is a more efficient trimming substrate (12, 33). To determine whether a similar relationship holds for the LTR pairwise strand transfer reaction, R33 and R55 were constructed. Two methods were used to compare the efficiencies of R35, R33, and R55 for strand transfer. First, identical, independent reactions were set up with R35, R33, or R55 as the substrate and efficiencies were compared. Second, a single reaction was run with equal amounts of the three substrates included, and then recombinants were subcloned and sequenced, and the ratio of R35, R33, and R55 recombinants was calculated.

A representative experiment done by the first method is shown in Table 1. In this experiment, R33 had a higher efficiency than R35, and R55 was below both. The trend of R33 being better than or equal to R35 and R55 being below both was observed for several experiments under various solution conditions.

Results from the second method are presented in Table 2. Two experiments were performed under identical conditions by using the optimal buffer for experiment 1 and a simplified buffer for experiment 2. Recombinants from the reactions were subcloned and sequenced. The results show that in both cases R33 and R35 appeared at a higher frequency than R55. Under optimal conditions, R33 was found at the highest frequency; with the simplified buffer, R35 predominated. We conclude that R33 is as efficient a substrate as, if not better than, R35. In addition, R55 was consistently integrated to the least extent.

Host site sequence analysis. During the present study, a significant number of integration sites were sequenced. The locations of the sites within the genome of lambda gtWES were mapped (Fig. 4A, sites shown for the major nonessential region). We plotted the number of integration events for each 1-kb segment of the region shown in Fig. 4A and compared this curve with the distribution of integration events from Fig. 3 of Brown et al. (1a) (Fig. 4B). For the set of AMV data, the sites appeared to be clustered around a peak in the 22- to 23-kb region and an approximately symmetrical decline in the regions bordering this segment. A similar trend was not immediately apparent in the smaller set of MoMLV data. It could be that the necessity of lambda lytic replication is responsible for the AMV clustering, but since there was little covariation between the AMV and MoMLV curves, such a restraint seems less likely. AMV IN appears to bind A-T-rich DNA preferentially (19), and a single-stranded DNA-RNA-binding activity has been demonstrated (21). To test for a possible correlation between A-T richness and integration activity, we plotted the percent

FIG. 4. Distribution of integration sites within lambda gtWES. (A) Integration sites for R35, R33, and R55 were plotted relative to the lambda gtWES genome. Note that lambda gtWES has a deletion and an inversion within this region relative to wild-type lambda. This map and that of Fig. 3, which is based on data from Brown et al. (1a), were constructed by taking these rearrangements into consideration (1). (B) The number of integrants for each 1-kb segment of the region depicted in panel A was plotted and compared with the distribution of the map of Fig. 3 (1a). Also plotted is the percent A-T for each segment relative to the lambda gtWES sequence.

A-T richness for each segment of the region depicted in Fig. 4A; note that this map is relative to lambda gtWES, which does not have a genome colinear with that of wild-type lambda. Although the region preferred by AMV has a higher A-T content than less efficiently used segments (63 versus 45%), the percent A-T richness does not have an exact correlation to the number of integration events.

To determine whether sequence bias existed at the immediate integration site, we aligned the sites relative to the centers of the host sequences duplicated. The sites were segregated by reaction conditions, LTR substrate, and size of duplication. The percent G-C richness was calculated for each position of the duplicated sequences, and the results are presented in Table 3. For the 6-bp duplications, a trend

towards G-C richness at the outside positions (those connected to the CA moieties) and A-T richness at the internal four positions was found (Table 3, R35 + R33 + R55). The average G-C content of the region depicted in Fig. 4 is 45.7%.

A recent report has been published which investigated the in vitro integration of MoMLV into DNA with and without nucleosomes (23). For naked DNA, a bias was found for A-T pairs at positions 2 bases away from the edges of the 4-bp duplication produced by MoMLV integration. We did a similar analysis for the set of 68 6-bp duplication sites (Fig. 5A). It appears that the sequence trend observed in Table 3 is part of a larger trend encompassing 14 bp around the duplication center (delineated by positions 7). At positions 3 bp away from the edge bases of the 6-bp duplicated sequences (positions 6; edge bases are marked by arrows), A-T richness of the magnitude found in the MoMLV study was apparent (>70% A+T). One base further on each side (positions 7) was G-C rich, producing a peak and trough with dyad symmetry about the duplication center. The high G-C contents of the edge bases (positions 3, arrows) were both followed by bases with high A-T contents (positions 2 and 4), producing a peak-and-trough pattern that was directly repeated, with no dyad symmetry about the duplication center.

Four additional alignments were performed: two using random numbers to generate sites within the region of Fig. 4 (Fig. 5B) and two that were near the 6-bp integration sites, with one aligned starting at position 16 3' from the duplication center and the other within the same region but at a variable distance from the duplication site (Fig. 5C). In all four, no symmetry characteristics similar to that of the integration alignment were found, and in general the values did not reach the extremes of the integration set. To determine whether the frequencies of G-C to A-T at each position were drawn from a homogeneous binomial distribution, a chi-square test was run on the five alignments; only the int:aligned curve of panel A differed significantly ($\chi^2_{29} = 95.46$; $P < 0.001$). The mean G-C contents for the 30 positions of the alignments of panels A and C are 40.5, 39.6, and 40.1% for int:aligned, near int:aligned, and near int: not aligned, respectively. For ran:dice and ran:number, the mean percents G+C are 46.2 and 45.6%, respectively. A t test showed that the means of the curves of panels A and C differ significantly from the expected random mean of 45.7% ($P = 0.014$, panel A, int:aligned; $P < 0.001$, panel C, near int:aligned and not aligned). Averaging the curves in panels B and C demonstrates graphically the trend of higher A-T composition near the integration sites (Fig. 5D).

TABLE 3. Host site analysis^a

Duplication size (bp)	Base pair positions	% G+C (no. of sites in set)			
		R35	R33	R55	R35 + R33 + R55
7	1-7	83, 67, 83, 67, 67, 67, 67 (6)	100, 50, 0, 0, 100, 50, 50 (2)	(0)	88, 63, 63, 50, 75, 63, 63 (8)
6	1-6	65, 20, 30, 40, 38, 65 (40)	44, 22, 44, 56, 33, 61 (18)	80, 30, 70, 20, 60, 70 (10)	62, 22, 40, 41, 40, 65 (68)
5	1-5	63, 38, 13, 38, 50 (8)	50, 50, 50, 50, 25 (4)	67, 100, 33, 17, 83 (6)	67, 61, 28, 33, 56 (18)

^a Integration events that produced host site duplications were 5 to 7 bp long. The events sequenced were categorized by duplication size and LTR substrate (R35, R33, and R55). The percent G+C richness for each position in the direct duplication was calculated. The outside positions are connected to the LTR ends, and for R35 the left position is that connected to the U5 end. The R35, R33, and R55 sets were compiled to form R35 + R33 + R55. The 6-bp duplications of R35 + R33 + R55 were used in the sequence analysis of Fig. 5.

FIG. 5. Analysis of host site sequence composition. (A) The 68 recombinants with 6-bp host sequence duplications were aligned relative to the center of the duplicated site (orientation of the LTR insert was taken into account such that the T7 end of the pGEM-3 constructs points toward the left of the graph). The percent G-C composition for each position of the 30 bases encompassing the host duplication site is graphed. The base pairs at the edges of the duplicated host sequences are marked with arrows (positions 3), and positions with symmetry relationships are numbered. (B) Two 30-bp alignments using random sites (68 sites each) within the region depicted in Fig. 4. are graphed. The random sites were generated by using either a random-number table (ran:number) or by rolling dice (ran:dice). (C) Two 30-bp alignments using the 68 sites of panel A but not centered around the insertion site. The sites of near int:aligned begin at position 16 3' to the integration site relative to the plus strand of the lambda genome and continue in that direction for 30 bp (orientation of the LTR insert was not taken into account). Near int: not aligned is from the same region but at a variable distance from the integration site. (D) The curves of panel B were averaged for each position (ave:ran), as were those of panel C (ave:near int).

DISCUSSION

We used an assay system which scores concerted LTR terminus integration catalyzed by AMV IN. AMV IN was the only viral protein included in the reactions, and thus understanding of changes in integration efficiency was simplified. By using this system, we were able to show, under optimal conditions, a pairwise integration efficiency that approached the activity provided by extracts of MoMLV-infected cells (1a). Additionally, the response of purified AMV IN to the solution parameters studied was similar to that of the integration activity contained in the cell extracts. That the response of the viral nucleoprotein complex within the cell extracts is similar to that of purified IN appears plausible, since additional studies have shown that human immunodeficiency virus IN is a major component of viral complexes isolated from human immunodeficiency virus-infected cells (11).

Although more complex nucleoprotein structures mediate integration and may be responsible for coupling of 3'-end processing and strand transfer, producing a constant duplication of the host site, only the immediate sequences of the LTR inverted repeat and IN are sufficient to confer integration potential upon a DNA molecule. For 3'-end processing, we have shown that AMV IN is more active on U3 substrates. In this study, a preference was again found for a U3 substrate over a U5 substrate in pairwise LTR strand transfer. That the relationship was the same for both steps of integration supports the possibility that one active site catalyzes both reactions. Emerging mutagenesis data also

indicate one active site, in that mutations that affect one activity, in general, affect the other activity (3, 16, 25, 30). Support for a single active site is advanced because viral DNA 3'-end processing and integration appear to occur by a common mechanism of direct nucleophilic attack (9, 32).

Because plasmid substrates were used to study the activities of the U3 and U5 sequences, it was possible to show that a synergistic relationship between the imperfect inverted repeat sequences does not affect catalysis of strand transfer; R33 was as efficient as, if not better than, R35. Why the U3 sequences are more efficient biochemically is not known; however, it may be that the U5 sequences of the avian viruses are constrained because of their effect on reverse transcription (6) in a manner similar to the proposal that human immunodeficiency virus U3 sequences are constrained from evolving to the most effective substrate for integration because they are also part of the *nef* open reading frame (3).

Selection of the host site during integration has been assumed to be random at the local sequence level on the basis of sequencing of a limited number of proviruses. However, recent studies which have sequenced larger numbers of recombinants (23), or total populations of LTR oligonucleotide integration products (20), indicate that the process is not totally random. We have found additional indications that host site selection is not random. When the 68 recombinants with 6-bp duplications were aligned relative to the duplication center, a pattern of high and low G-C contents at positions symmetrically related about the dupli-

cation center was observed. This pattern deviates significantly from that expected for a random-selection process. The pattern of sequence composition appeared to be confined to 14 bp around the duplication site, but it is of interest that the sequence composition immediately around the site was significantly more A-T rich than that found in the random alignments. On a larger scale, the integration events seemed to be clustered and the preferred regions had higher A-T contents. A-T-rich segments of DNA may recruit integrase in a general manner to a DNA target, with determination of the specific sequence where strand transfer occurs by sequence characteristics which promote nucleophilic attack of the phosphate backbone by AMV IN. The G-C content pattern seen in Fig. 5A may be an indication of sequence patterns that give rise to conformational states that are preferred as host sites.

ACKNOWLEDGMENTS

We thank C. Hall for typing the manuscript. Statistical analysis was performed by George Vogler, Division of Biostatistics, Washington University Medical Center.

This work was supported by grant CA16312 from the National Cancer Institute.

REFERENCES

- Brown, P. O. Personal communication.
- Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop. 1987. Correct integration of retroviral DNA *in vitro*. *Cell* 49:347-356.
- Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop. 1989. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc. Natl. Acad. Sci. USA* 86:2525-2529.
- Bushman, F. D., and R. Craigie. 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.
- Bushman, F. D., T. Fujiwara, and R. Craigie. 1990. Retroviral DNA integration directed by HIV integration protein *in vitro*. *Science* 249:1555-1558.
- Chow, S. A., K. A. Vincent, V. Ellison, and P. O. Brown. 1992. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science* 255:723-725.
- Cobrinck, D., A. Aiyar, Z. Ge, M. Katzman, H. Huang, and J. Leis. 1991. Overlapping retrovirus U5 sequence elements are required for efficient integration and initiation of reverse transcription. *J. Virol.* 65:3864-3872.
- 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.
- Donehower, L. A., and H. E. Varmus. 1984. A mutant murine leukemia virus with a single missense codon in *pol* is defective in a function affecting integration. *Proc. Natl. Acad. Sci. USA* 81:6461-6465.
- Engelman, A., K. Mizuuchi, and R. Craigie. 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* 67:1211-1221.
- Farnet, C. M., and W. A. Haseltine. 1990. Integration of human immunodeficiency virus type 1 DNA *in vitro*. *Proc. Natl. Acad. Sci. USA* 87:4164-4168.
- Farnet, C. M., and W. A. Haseltine. 1991. Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. *J. Virol.* 65:1910-1915.
- Fitzgerald, M. L., A. C. Vora, and D. P. Grandgenett. 1991. Development of an acid-soluble assay for measuring retrovirus integrase 3'-OH terminal nuclease activity. *Anal. Biochem.* 196:19-23.
- 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* 86:3065-3069.
- Fujiwara, T., and K. Mizuuchi. 1988. Retroviral DNA integration: structure of an integration intermediate. *Cell* 54:497-504.
- Grandgenett, D. P., A. C. Vora, and R. D. Schiff. 1978. A 32,000-dalton nucleic acid-binding protein from avian retrovirus cores possesses DNA endonuclease activity. *Virology* 89:119-132.
- Kahn, E., J. P. G. Mack, R. A. Katz, J. Kulkosky, and A. M. Skalka. 1990. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Res.* 19:851-860.
- Katz, R. A., G. Merkel, J. Kulkosky, J. Leis, and A. M. Skalka. 1990. The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. *Cell* 63:87-95.
- Katzman, M., R. A. Katz, A. M. Skalka, and J. Leis. 1989. The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the *in vivo* sites of integration. *J. Virol.* 63:5319-5327.
- Knaus, R. J., P. J. Hippenmeyer, T. K. Misra, D. P. Grandgenett, V. R. Muller, and W. M. Fitch. 1984. Avian retrovirus pp32 DNA binding protein: preferential binding to the promoter region of long terminal repeat DNA. *Biochemistry* 23:350-359.
- 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.
- Mumm, S. R., and D. P. Grandgenett. 1991. Defining nucleic acid-binding properties of avian retrovirus integrase by deletion analysis. *J. Virol.* 65:1160-1167.
- Panganiban, A. T., and H. M. Temin. 1984. The retrovirus *pol* gene encodes a product required for DNA integration: identification of a retrovirus *int* locus. *Proc. Natl. Acad. Sci. USA* 81:7885-7889.
- Pryciak, P. M., A. Sil, and H. E. Varmus. 1992. Retroviral integration into minichromosomes *in vitro*. *EMBO J.* 11:291-303.
- Quinn, T. P., and D. P. Grandgenett. 1988. Genetic evidence that the avian retrovirus DNA endonuclease domain of *pol* is necessary for viral integration. *J. Virol.* 62:2307-2312.
- Roth, M. J., P. Schwartzberg, N. Tanese, and S. P. Goff. 1990. Analysis of mutations in the integration function of Moloney murine leukemia virus: effect on DNA binding and cutting. *J. Virol.* 64:4709-4717.
- Roth, M. J., P. L. Schwartzberg, and S. P. Goff. 1989. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell* 58:47-54.
- Schwartzberg, P., J. Colicelli, and S. P. Goff. 1984. Construction and analysis of deletion mutations in the *pol* gene of Moloney murine leukemia virus: a new viral function required for establishment of the integrated provirus. *Cell* 37:1043-1052.
- Silhavy, T., M. Berman, and L. Enquist. 1984. Experiments with gene fusions, p. 140. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- Varmus, H. E. 1983. Retroviruses, p. 411-503. *In* J. A. Shapiro, ed., *Mobile genetic Elements*. Academic Press, Inc., New York.
- Vink, C., D. C. van Gent, Y. Elgersma, and R. H. A. Plasterk. 1991. Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage. *J. Virol.* 65:4636-4644.
- Vink, C., D. C. van Gent, and R. H. A. Plasterk. 1990. Integration of human immunodeficiency virus types [sic] 1 and 2 DNA *in vitro* by cytoplasmic extracts of Moloney murine leukemia virus-infected mouse NIH 3T3 cells. *J. Virol.* 64:5219-5222.
- Vink, C., E. Yeheskiely, G. A. van de Marel, J. H. van Boom, and R. H. A. Plasterk. 1991. Site-specific hydrolysis and alcoholysis of human immunodeficiency virus DNA termini mediated by the viral integrase protein. *Nucleic Acids Res.* 19:6691-6698.
- Vora, A. C., M. L. Fitzgerald, and D. P. 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.