

## Intramolecular Integration Within Moloney Murine Leukemia Virus DNA

CHARLES SHOEMAKER, JOSEPH HOFFMANN, STEPHEN P. GOFF, AND DAVID BALTIMORE\*

*Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology,  
Cambridge, Massachusetts 02139*

Received 10 April 1981/Accepted 11 June 1981

By screening a library of unintegrated, circular Moloney murine leukemia virus (M-MuLV) DNA cloned in  $\lambda$  phage, we found that approximately 20% of the M-MuLV DNA inserts contained internal sequence deletions or inversions. Restriction enzyme mapping demonstrated that the deleted segments frequently abutted a long terminal repeat (LTR) sequence, whereas the inverted segments were usually flanked by LTR sequences, suggesting that many of the variants arose as a consequence of M-MuLV DNA molecules integrating within their own DNA. Nucleotide sequencing also suggested that most of the variant inserts were generated by autointegration. One of the recombinant M-MuLV DNA inserts contained a large inverted repeat of a unique M-MuLV sequence abutting an LTR. This molecule was shown by nucleotide sequencing to have arisen by an M-MuLV DNA molecule integrating within a second M-MuLV DNA molecule before cloning. The autointegrated M-MuLV DNA had generally lost two base pairs from the LTR sequence at each junction with target site DNA, whereas a four-base-pair direct repeat of target site DNA flanked the integrated viral DNA. Nucleotide sequencing of preintegration target site DNA showed that this four-base-pair direct repeat was present only once before integration and was thus reiterated by the integration event. The results obtained from the autointegrated clones were supported by nucleotide sequencing of the host-virus junction of two cloned M-MuLV integrated proviruses obtained from infected rat cells. Detailed analysis of the different unique target site sequences revealed no obvious common features.

During infection of susceptible cells by a retrovirus, circular DNA reverse transcripts of viral RNA molecules accumulate in the cell nucleus. Several lines of evidence suggest that circular DNA may be the form that integrates into cellular chromosomes and generates the integrated proviral DNA (12, 21, 32). Most of the circular DNA molecules have one of two forms: either they have two long terminal repeat (LTR) sequences as if the linear reverse transcripts were circularized by blunt-end ligation, or they have a single LTR sequence as if they arose by homologous recombination of the LTR ends of linear DNA (20, 33).

The exact mechanism of integration is uncertain, but two hallmarks of the process have been observed: at the site of integration two bases of the LTR sequence are usually lost, and four to six bases of the chromosomal DNA become reiterated (4, 15, 24, 26). The integrated DNA is similar in structure to DNA sequences that constitute the transposable DNA elements of bacteria (reviewed in reference 13), the *copia* and

related sequences in *Drosophila* (5), and the *Ty1* elements of yeasts (7, 8).

We recently observed that some of the circular retrovirus DNA molecules have unusual structures, and we interpreted one such structure as a consequence of unintegrated retrovirus DNA integrating into itself (26). Here, we extend that observation to demonstrate that deleted and reorganized circular DNA molecules occur at a high frequency in Moloney murine leukemia virus (M-MuLV)-infected cells. The structure of most of these DNA molecules suggests that the LTR sequence is the site of a variety of recombinational events that can lead to either viral integration into chromosomes or reorganized circular DNA molecules.

### MATERIALS AND METHODS

**Phage library construction.** The construction of a Charon 21A phage (2) library containing unintegrated circular M-MuLV DNA inserts has been previously described (26).

Rat NRK-5 cell DNA was prepared as described by

Steffen and Weinberg (28). After complete digestion of the DNA with *Eco*RI the DNA was resolved by electrophoresis through a 0.6% agarose gel. The region of the gel containing DNA of 8 to 23 kilobases was removed, and the DNA was purified with glass powder by the procedure of Vogelstein and Gillespie (31). Charon 4A DNA was digested with *Eco*RI, and the annealed end fragments were prepared and ligated to the NRK-5 DNA by the method of Maniatis et al. (16). The ligated DNA was packaged (29), plated onto LE392 cells, and screened for hybridization to  $^{32}$ P-labeled M-MuLV DNA by the procedure of Benton and Davis (1). Phage containing M-MuLV sequences were plaque purified, and the recombinant inserts were transferred to pBR322 at the *Eco*RI site.

**Phage DNA preparation.** Large-scale phage DNA preparations were performed as described by Enquist et al. (6). It was frequently necessary to prepare only the small amount of phage DNA obtained from a single plate stock of phage. This was done by a modification of the procedure of Cameron (Ph.D. Thesis, Stanford University, Stanford, Calif., 1976) as follows. Approximately  $10^5$  to  $10^6$  phage were absorbed to LE392 cells and plated onto 1% agarose (Sigma Chemical Co.) in a 10-cm petri dish containing a nutrient broth and incubated for 8 to 15 h at 37°C. Five milliliters of 10 mM Tris-10 mM MgCl<sub>2</sub> (pH 7.5) and several drops of chloroform were added to the plates, and the plates were allowed to sit overnight at 4°C. The liquid (clear lysate) was removed to a sterile tube containing a few drops of chloroform. To a 1.5-ml Eppendorf microfuge tube was added 400  $\mu$ l of clear lysate, 50  $\mu$ l of 2 M Tris-0.2 M EDTA (pH 8.5), and 15  $\mu$ l of 20% sodium lauryl sulfate. The mixture was blended in a Vortex mixer and incubated at 65°C for 15 min. Then, 50  $\mu$ l of 5 M potassium acetate was added, and the mixture was left on ice for 30 min. The tube was then centrifuged for 15 min in an Eppendorf centrifuge, and the supernatant was extracted once with phenol and twice with chloroform. One milliliter of ethanol (-20°C) was added, and the Eppendorf tube was immediately centrifuged for 5 min at 4°C. The pellet was dissolved in 400  $\mu$ l of 0.3 M sodium acetate-2  $\mu$ g of RNase per ml and incubated at room temperature for 15 min. One milliliter of ethanol was added, and the solution was frozen in dry ice and then centrifuged for 15 min at 4°C. The DNA pellet was washed with ethanol, dried in vacuo, and dissolved in 50  $\mu$ l of 10 mM Tris-1 mM EDTA.

**Analysis of recombinant DNA.** Restriction enzyme digestion was performed as recommended by the supplier (New England Biolabs). Gel transfer and filter hybridization was done by the method of Southern (27). The Maxam and Gilbert (17) method of DNA sequencing was employed.

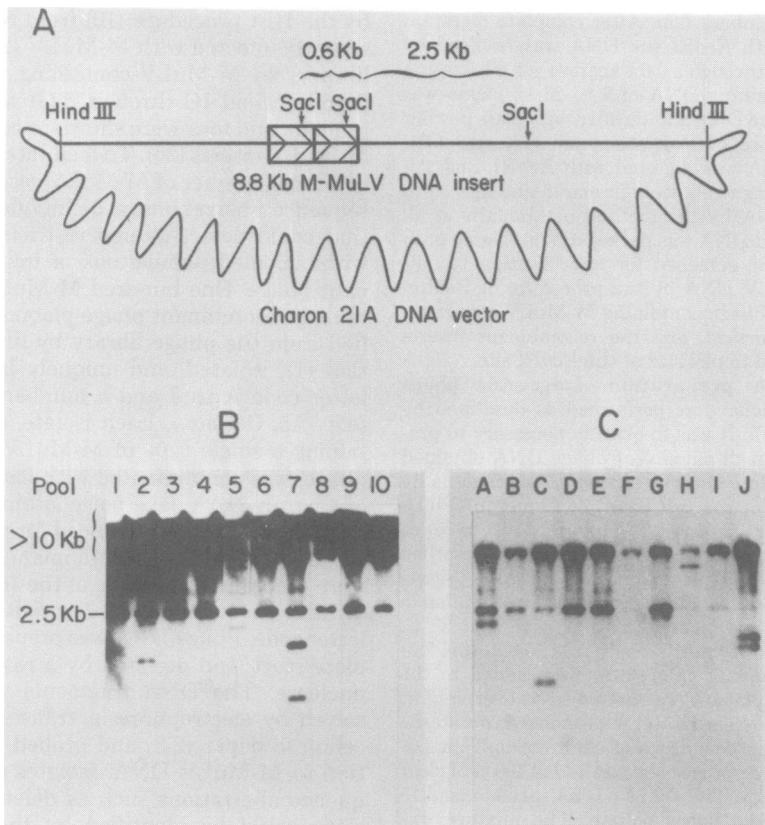
## RESULTS

**Screening a phage library of unintegrated M-MuLV DNA for variant inserts.** We have previously reported the construction of a Charon 21A phage library containing recombinant inserts at the *Hind*III site that were derived from the small, circular DNA isolated

by the Hirt procedure (10) from NIH/3T3 cells recently infected with M-MuLV (26). From this library, 20 M-MuLV-containing recombinants (herein called 1G through 20G) were picked at random, and four were shown to contain variant M-MuLV inserts (26). To facilitate the screening of a large number of M-MuLV recombinants for sequence aberrations, a technique was devised that could detect altered restriction fragments within pooled populations of impure recombinant phage. One hundred M-MuLV DNA-containing recombinant phage plaques were identified from the phage library by filter hybridization (1), isolated, and uniquely labeled with a letter code A to J and a number code 1 to 10 (e.g., A5, G8, etc.). Each isolate, although containing a single type of M-MuLV recombinant phage, was contaminated with various amounts of Charon 21A phage not containing M-MuLV sequences. To avoid the need for plaque-purifying each M-MuLV recombinant, plate stocks were prepared from pools of the 10 isolates having either a common number code or a common letter code. Phage DNA was prepared from each plate stock and digested by a restriction endonuclease. The DNA fragments were then resolved by electrophoresis, transferred to nitrocellulose paper (27), and probed for hybridization to M-MuLV DNA. Isolates containing sequence aberrations such as deletions or inversions could be identified by the presence of altered restriction fragments detected among the two plate stock phage pools that are unique to each isolate (see below).

The first screen was performed by digesting phage pools with *Sac*I. This restriction enzyme digests wild-type, 8.8-kilobase M-MuLV DNA at a site within the LTR and at an additional site within unique M-MuLV sequence (Fig. 1A). Because Charon 21A DNA has no *Sac*I restriction sites, the only M-MuLV DNA-containing restriction fragments that will resolve on 1.5% agarose gels are the 600-base-pair LTR fragments and an internal 2.5-kilobase fragment (Fig. 1A). Deletions of sequence between the *Sac*I sites or sequence inversions that alter the relative locations of those sites will result in *Sac*I fragments of altered size.

When the ten plate stock phage pools having a common number code were analyzed (Fig. 1B), six variant fragments were observed—two each in pools 2, 5, and 7 (the faint band seen above the 2.5-kilobase band in some pools is probably a 3.1-kilobase partial digest product). Plate stock phage pools were then prepared containing the three isolates having a common letter code and the number 2, 5, or 7. The same variant *Sac*I fragments could again be detected, thus identify-



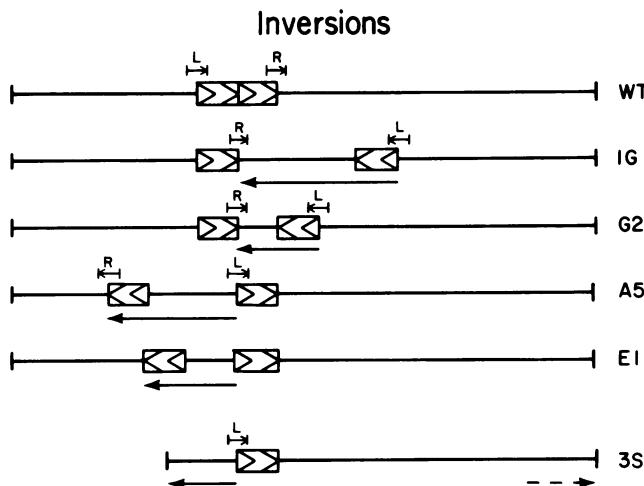
**FIG. 1. Large-scale screening of M-MuLV recombinant inserts for gross sequence aberrations.** *A*, Diagrammatic representation of a wild-type 8.8-kilobase M-MuLV recombinant phage. LTR sequences are boxed, and the 5' to 3' direction is left to right for the DNA strand with the same polarity as viral RNA. *B*, Plate stock pools were prepared as described in the text. Phage DNA was isolated and digested with *SacI*. The DNA fragments were resolved on 1.5% agarose gels, transferred to nitrocellulose filters, and probed for M-MuLV DNA sequences. An autoradiogram of the filters prepared from plate stock pools 1 through 10 is shown. *C*, Autoradiogram of filters prepared as above from plate stock pools *A* through *J*. These pools contained DNA from numbers 2, 5, and 7 only.

ing the specific phage isolates containing variant M-MuLV recombinant inserts as A5, C7, G2, H5, J2, and J7.

A second screen was performed by digesting phage pools with *HindIII*. Because *HindIII* was the enzyme employed for the original cloning to form the phage library, this enzyme excised the entire recombinant inserts from the phage. Inserts containing deletions were thus identified by their increased electrophoretic mobility. The screening process was similar to that used for *SacI* digests, except that a 0.7% agarose gel was used to resolve the DNA fragments. Due to the presence of a large number of deletion variants it was necessary to construct 20 phage pools, each containing the 10 isolates with a common number or common letter code. This screening process (data not shown) allowed the identification of 13 additional M-MuLV DNA variant inserts.

After identification, either the variants were plaque purified for direct mapping of restriction endonuclease sites or the DNA was mapped without purification by the procedure of Southern (27). Making use of the extensive and well-established restriction map for wild-type M-MuLV (9), it was possible to identify the locations of the deletion or inversion endpoints to within about 300 base pairs in every instance. The results of the restriction enzyme mapping are shown in Fig. 2 for variants containing inverted segments and in Fig. 3 for variants containing deletions. A total of 23 variants were identified by the combined screening of 120 M-MuLV-containing recombinant phage. Thus, we can estimate that approximately 20% of the M-MuLV inserts in our library contain either deleted or inverted segments.

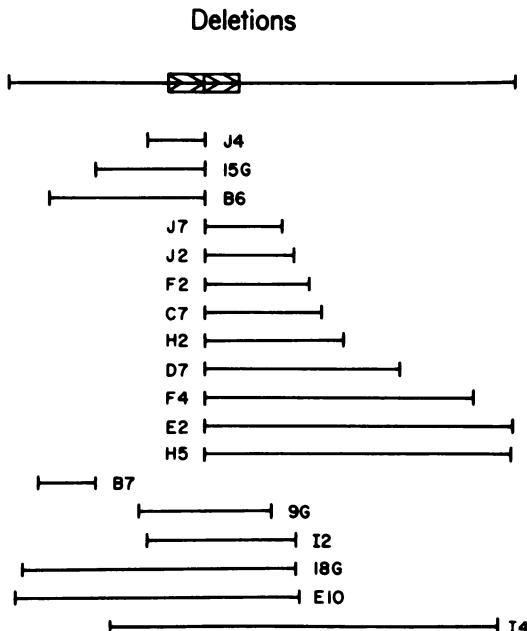
Many of the variant inserts contained alterations with an endpoint very close to the terminus



**FIG. 2. Diagrammatic maps of M-MuLV recombinant DNA inserts containing sequence inversions.** The top diagram labeled WT is a wild-type 8.8-kilobase M-MuLV insert as shown in Fig. 1. Solid arrows under each variant insert indicate the region of DNA which is inverted. Bracketed arrows designate the regions within each clone that was sequenced and presented in Fig. 4. L is a left-end sequence, and R is a right-end sequence. The direction of the arrow indicates the right-to-left presentation of that sequence in Fig. 4. The dashed arrow under 3S indicates the region identical to the inverted segment in that clone.

of an LTR sequence. Because the LTR termini are known to be sites of recombination with host DNA during retrovirus integration (3, 11, 19), it seemed likely that these variants arose by integrative processes in a manner similar to that described in an earlier report (26). To analyze more precisely these variant recombinant inserts, their nucleotide sequence was determined near the positions of variation from wild-type structure ("junction sequences"). Figure 4 presents the junction sequences of many of these variants as well as the sequences across the host-virus junction of two cloned integrated proviruses (ZIP and ZAP, described below). Where the left end of an LTR sequence (as oriented in Fig. 1) is involved, the "left-end sequence" in Fig. 4 presents the sequence 5' to 3' across the deleted segment, inversion endpoint, or host-virus junction, reading into the LTR sequence. Similarly, where the right end of an LTR sequence is involved, the "right-end sequence" is presented 5' to 3' in Fig. 4 from within the LTR sequence and across the recombined region.

**Integrative inversions.** There were four sequence inversions analyzed that involved no accompanying deletion and one (3S) that was both inverted and deleted (3S was isolated by an independent screening of the Charon 21A phage library described earlier). The junction sequences from three of the four simple inversions were determined. In addition, the nucleotide sequence was obtained from a wild-type M-MuLV DNA clone at the sites at which junction sequences occurred (Fig. 4; inversion 1G has



**FIG. 3. Locations of internal sequence deletions within M-MuLV recombinant DNA inserts.** At the top is a wild-type 8.8-kilobase M-MuLV insert as shown in Fig. 1. The lines underneath indicate the regions deleted within the designated deletion-containing inserts.

been described previously by Shoemaker et al. [26]). In each case, the inversion junctions involved ends of LTR sequences from which the two terminal A-T base pairs had been lost. Four

## **LEFT END SEQUENCES**

	-30	-25	-20	-15	-10	-5	$\alpha\beta\gamma\delta$	left LTR sequence
M-MuLV	AATAAA	G	TTTTT	TTT	AGT	CTCCAGAAAAAGGGGGC	GAA	TGAAAGACCCCACCTGTAGGTTGCCAAGCTAGC
ZIP	CTGTCAATTCGTTGATCTT	T	TTT	GGTAC	CTACATG			
ZAP	TCC	CTG	CTGCCAGT	CTGCCCGCCCCG	GGGTAC	CTGG		
1G	CTACGGCGTT	CTT	CTT	CTGTTCTCT	CTCTGA			
G2	CCATGATGAGGAGGG	T	CC	CAGTAAGCT	TACCGGGTTG			
A5	AAGACCTTGATCTTAA	AC	CTTAGGTGATGAGGTCT	GTT				
3S	ATAACGGCTATCAGT	TAA	ACATTAGCTT	ACCTT				
15G	CGTTTGTATGTC	CCGGACA	AGATCC	CAGGGCTTACATT				

## **RIGHT END SEQUENCES**

	right LTR sequence	$\alpha\beta\gamma\delta$	5	10	15	20	25	30
M-MuLV	GAGTGTATTGACTACCCGTCAAGGGGGCTTTCA	T	TTGGGGCTCGTCCGGATCGGAGACCCCTGCCAG					
ZIP		CATCCAGCACGTATCAAATTAAATTGGTTTTTCT						
ZAP		CTGGGAGAACCTTCACTGAACAACCAAAAGTGACCTGCA						
IG		CTGATACTGTCTCTCTTCTCCGGGTTCTCGGT						
G2		GTGGTCACAGCCCCAATAGGCACAGTAGAAGGAGTC						
A5		GGTAAAGGTGCCGTCTCGCGGCCATCCGACGTTAAAG						
J2		AAGACCACCCCCCTCCGACAGGAGGAAATGGTGAGA						
C7		CCACCTAGTCTCACTCTGCCACTTGTCCAGGGCT						
H5		( )GATCAGAACACTCAAATATCACTGAGACCCAAAGC						
J7	( )	)AGCCCTTGTACACCCCTAAGCTCCGGCTCTCTCT						

## TARGET SITE SEQUENCES

## **CONSENSUS SEQUENCE**

-30 -25 -20 -15 -10 -5 αβγδε 5 10 15 20 25 30  
 .T.....T.....CT.C.GT.....TCTAC...C.....CC..C..A.....T.....

**FIG. 4.** Relevant nucleotide sequences obtained from *M-MuLV* wild-type and variant recombinant inserts and integrated proviruses. Shown are the junction sequences surrounding the endpoints of inversions, deletions, and integrations of *M-MuLV* DNA. Each sequence is identified by the clone number from which it was derived—ZIP and ZAP refer to cloned, integrated *M-MuLV* DNA derived from infected rat cells, and *M-MuLV* refers to the sequence of a wild-type 8.8-kilobase *M-MuLV* recombinant DNA insert. Left-end sequences and right-end sequences refer to the sequences surrounding the left or right LTR termini in which the LTR is oriented as shown in Fig. 1. The sequences are presented (only the 5'-to-3' strand is shown) as being representative of integration events, so that left-end sequences read from the target site into the left LTR, and the right-end sequences read from the LTR into the target site sequences. Because all LTR sequences were identical, they are presented as straight lines extending over the actual region that was sequenced. Parentheses enclose bases that were absent in two of the deletion clones. The four repeated target site bases are labeled  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ; for the *M-MuLV* sequence, these positions include the two bases of LTR sequence lost from each inversion, deletion, or integration endpoint and are indicated by a box. Target site sequences are the preintegration sequences surrounding the integration target sites. Except for ZIP and ZAP, these sequences were obtained by sequencing wild-type *M-MuLV* cloned inserts. The target site sequences for ZIP and ZAP were extrapolated from the left-end and right-end sequences by assuming that the four-base-pair direct repeats at the host-virus junctions existed only once before integration. The sequence to the 5' side of the four-base reiteration is designated by negative numbers indicating the distance in base pairs from  $\alpha$ . The sequence to the 3' side of the four-base reiteration is designated by positive numbers indicating the distance in base pairs from  $\delta$ . The consensus sequence presents bases from the target site sequence which were found at a specified position in 50% or more of the cases. The circled base within *M-MuLV* unique sequence at position -27 was shown to have been mutated in one variant insert (20G) to a cytosine residue generating a novel HindIII site.

base pairs present only once in the wild-type sequence were found at both inversion endpoints (Fig. 4). These inversions were thus congruent in general structure to the previously described inversion.

Inversion 3S was unique among the variants in that its inverted segment was not bounded by LTRs. As indicated by restriction mapping and later confirmed by sequence analysis (data not shown), this clone contained an exact inverted repeat of 0.8 kilobase pairs at each end. The region at the left end of clone 3S (as oriented in Fig. 2) between the *Hind*III cloning site and the LTR contained, in inverted orientation, DNA that was identical in sequence to that found between 4.35 and 5.15 kilobase pairs on wild-type M-MuLV DNA (9).

**Integrative deletions.** Of the 18 deletion-containing M-MuLV DNA inserts that were mapped by restriction enzyme digestion, 12 occurred with one end near an LTR terminus. Five of these (15G, J2, C7, H5, and J7) were sequenced to locate the precise position of the deletion. For three of these (15G, J2, and C7), one edge of the deleted segment began two base pairs within the LTR sequence (Fig. 4). The deleted segment within the H5 clone included three base pairs of the LTR. In J7, the deletion included 23 base pairs of the LTR and the adjacent unique M-MuLV DNA. Six of the 18 deletions were not near LTR termini, and 5 of these completely deleted all of the LTR sequence from the insert (Fig. 3).

**Host-virus junction sequences of M-MuLV proviruses.** A phage library (Charon 4A) was constructed from *Eco*RI-digested DNA prepared from NRK-5 cells, a cloned rat cell line containing six M-MuLV proviruses (28). Recombinant phage containing M-MuLV sequences were isolated and plaque purified, and the cloned inserts were transferred to pBR322 (J. Hoffmann, J. Gusella, C. Tabin, and R. Weinberg, unpublished data). The two distinct DNA inserts isolated in this way, ZIP (a clone that did not generate infectious virus after transfection onto NIH/3T3 cells) and ZAP (an infectious clone), were sequenced across the junction between the rat host cellular DNA and the M-MuLV provirus (Fig. 4). Two base pairs from the LTR termini were absent at the host-virus junction, and a four-base-pair direct repeat of host DNA sequences flanked the provirus.

**Specificity of M-MuLV for the target site sequence.** As discussed below, the structure of all of the inserts in which the sequence was deleted near an LTR terminus, or where inversions occurred, is consistent with their generation by an autointegrative process. The sequence abutting the LTR within each of these junction

sequences therefore represents a target site for an M-MuLV integration event. Thus, this series of variants provides a library of target sites from which to search for regularities. Figure 4 shows the nucleotide sequences of eight preintegration target sites. The target sites for ZIP and ZAP were extrapolated from the sequence obtained at the host-LTR junctions by assuming that the four-base-pair reiteration at the junctions existed only once before integration. The remaining six target sites came from direct sequencing of the relevant portions of a cloned wild-type M-MuLV DNA insert.

Most of the target site sequences were analyzed for common features by computer analysis (18). No significant sequence homologies could be detected which were common to more than a small subset of the total target sites. This interpretation was true even when each strand was compared with the complementary strands from the other target sites, thus assuming no specific orientation of integration. (We are grateful to Cary Queen for performing this analysis.)

A second method of examining the target sites for common features was to determine the frequency of occurrence of specific nucleotides as a function of their position relative to the precise integration site. In Fig. 4 a consensus sequence is shown presenting the nucleotides which were found to have been present at specific locations (relative to the integration site) at a frequency 50% or greater. (Included in the consensus sequence are target site sequences from an additional M-MuLV variant insert [19G] which contains an integrative deletion and the target site sequence extrapolated from a report by Dhar et al. [4].) If there were a completely random sequence within our target site survey we would have expected 16 of the 72 target site base positions to fit this criterion. Thus, the finding of 18 positions within the consensus sequence is not particularly significant. However, several features of the consensus sequence suggest that there may be a low level of target site specificity. The frequently occurring bases shown in Fig. 4 seemed to cluster within three regions, most obviously between -4 and  $\alpha$ , but also from -18 to -12 and from 7 to 14. Also suggestive was the observation that base position -2 (on the DNA strand shown) contained a pyrimidine residue in all 10 of the sequences examined (7 out of 10 bases in that position were T), and that between base positions 6 and 12 we found 56 pyrimidine residues out of 77 examined including 35 cytosine residues. It is interesting to note that this C-rich region was almost directly opposite a stretch of G residues within the right end of the LTR sequence.

**Reverse transcriptase errors in vivo.** In

the course of this analysis, we had occasion to sequence a total of approximately 900 base pairs of DNA that were obtained from the same region of two independently isolated clones. No instance of a single-base-pair alteration was found. The M-MuLV recombinant clones were generated after infection with virus prepared from a producer cell line that was biologically cloned through single cell-single virus techniques (23). Several rounds of viral infection could have occurred during the production of the M-MuLV producer cell line and also after the viral infection which led to the cloned DNA inserts. Therefore, several reverse transcription cycles probably took place between the viral RNA that gave rise to the producer cell and the cloning of the M-MuLV reverse transcripts. Thus, the significance of finding no sequence divergence over 900 base pairs is amplified.

Although no reverse transcription errors were detected directly by sequencing, one apparent deletion-containing M-MuLV insert that was isolated contained, instead of a deletion, a new *Hind*III restriction site. The T-to-C point mutation (circled base, Fig. 4) resulted in the cloning of an insert containing only the portion of the M-MuLV genome between -0.03 and 5.15 kilobase pairs (9).

## DISCUSSION

In this report we have analyzed 120 M-MuLV DNA-containing recombinant phage, prepared from unintegrated circular DNA, for the presence of gross sequence aberrations. The data indicate that at a 24 h postinfection, approximately 20% of the unintegrated circular DNA molecules contain either deleted or inverted segments.

The origin of these aberrant forms of DNA is evident from the occurrence in them of the hallmarks of the retrovirus integration process. As discussed in a previous report (26) and as described by others, the hallmarks of retrovirus integration appear to be integration at the border of an LTR, deletion of the two terminal nucleotides from the LTR, and reiteration of a small number of nucleotides at the target site on either end of the integrated DNA (4, 15, 24). The data presented here on two integrated M-MuLV proviruses confirm these regularities and show that M-MuLV proviral DNA has a four-base repeat on either end as seen previously for the related Moloney sarcoma virus (4).

All of the clones with inverted segments that were analyzed had one inversion endpoint at the end of an LTR, had lost two bases of the LTR sequence, and had a four-base reiteration at either end of the inversion. Further, it was

proved that the reiterated four bases were present only once prior to the inversion event. Of the 18 deletions, 10 occurred near LTR borders, with 3 of 5 tested having a two-base deletion of LTR sequence. Thus, it seems evident that the inversion and many deletions were the consequence of an integrative process. Therefore, retrovirus DNA has a propensity to integrate into itself, generating either inversions or deletions (the consequences of autointegration should include both inversion and deletion depending on the orientation of the DNA strands during the event [22]).

It is not yet clear which form of retrovirus DNA is the direct precursor to integration. Evidence that form I circular DNA constitutes a major portion of unintegrated retrovirus DNA in the nucleus and is formed from a cytoplasmic linear DNA precursor (21) suggests that circular DNA may be the precursor to integration. Further evidence for this view comes from the observation that the *Fv-1* gene product, which blocks retrovirus infection, somehow prevents the circularization of linear retrovirus reverse transcripts (12, 32). The possibility can not be eliminated that linear DNA is the pre-integrative form and that the *Fv-1* gene product damages the linear DNA in a way which not only prevents its integration but also independently blocks circularization. Swanstrom et al. (30) report an apparent deletion of an avian sarcoma virus DNA reverse transcript which deleted only a small portion of the second LTR. It is unlikely for topological reasons that such a structure was generated by integration within a monomeric circle (although autointegration within a dimer circle could generate this structure); thus, this result suggests that linear retrovirus DNA may be a precursor to integration.

The deletions that did not terminate at LTR ends may not have occurred by integrative processes. It is possibly significant, however, that most of these putative nonintegrative deletions removed all of the LTR sequence, suggesting that the LTR might be able to catalyze deletion of itself along with the sequence from both sides. The observed deletions, however, may simply be random ones that have been selected for the retention of their *Hind*III restriction site.

Two of the putative autointegrative deletion variants that were isolated (in which the deletion endpoints mapped to the LTR ends) were lacking more than the two base pairs. In one case, deletion H5 (Fig. 4), three bases from the LTR terminus were deleted. In another, deletion J7, 23 base pairs from the LTR terminus were lacking (Fig. 4). The significance of these variant structures is not apparent at present, but they

suggest that the integration process may not invariably cause deletion of only two bases.

The most unusual variant was 3S, in which a duplicated segment of viral DNA was found in inverted orientation at the ends of a deleted segment of the DNA. A model to explain the origin of this recombinant phage insert is presented in Fig. 5. It assumes the existence of a dimer circle of M-MuLV DNA molecules in a head-to-tail orientation (dimers of this type have been identified in avian sarcoma virus-infected cells by Kung et al. [14]). Integration of one of the M-MuLV LTRs into unique sequence DNA of the second M-MuLV copy could give rise to a dimer circle containing an inverted segment. Upon *Hind* III digestion and cloning, recombinant inserts analogous to 3S will then arise. Although less likely, a structure such as 3S may have arisen by the integration of one unintegrated monomer M-MuLV DNA molecule within a second identical molecule or through integration of a linear M-MuLV dimer within itself.

The recombination event which generated the

3S variant had to have occurred before *Hind* III digestion of the DNA and cloning into the Charon 21A vector. Therefore, at least in the case of 3S, the integrative recombination does not occur during passage of the recombinant phage in prokaryotes. This provides additional evidence beyond that previously presented (26) that the variant molecules arise in the infected NIH/3T3 cell and not after molecular cloning.

Our results imply that M-MuLV consistently generates a four-base reiteration of host sequences upon integration while generally losing two base pairs from each LTR. This is the case for M-MuLV or the related murine sarcoma virus whether the target cell is murine (26), rat (ZIP and ZAP, this paper), or mink (4). Since mouse mammary tumor virus generates a six-base-pair repeat in rat cells (15), the data now seem to indicate that the target site sequence reiteration is specified by the virus rather than by the host. This is not yet unambiguously clear, though, as ZIP and ZAP could be five- and six-base-pair reiterations, respectively, in the unlikely event that three and four base pairs, respectively, were lost from the right LTR during integration.

The primary purpose for sequencing a large number of M-MuLV integration target sites was to analyze the sequences for common features. The most striking conclusion from these results was the lack of similarity between the various target sites. Computer analysis revealed no significant sequence homologies common to even a majority of these sequences. A search for some degree of base preference at specific sites within the target site also suggested at best only a very low level of target site specificity. Therefore, our results are essentially consistent with those of Shimotohno and Temin (25), who found no apparent target site specificity for spleen necrosis virus integration.

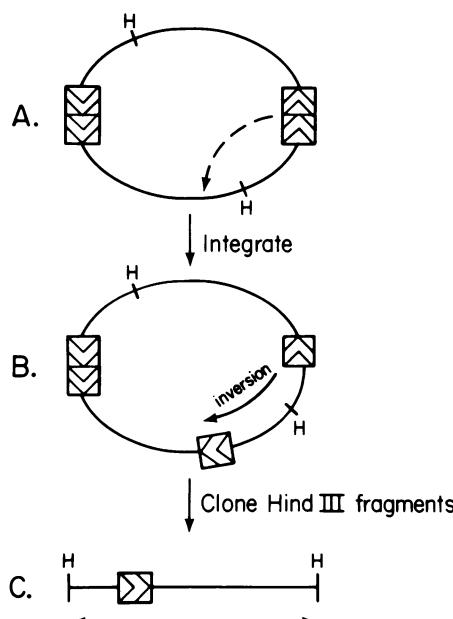
#### ACKNOWLEDGMENTS

We thank Michael Paskind for assistance with the DNA sequencing and Eli Gilboa for frequent and fruitful discussion. We are also grateful to Cary Queen for performing the computer analysis comparing the target site sequences. Finally, we thank Cliff Tabin for testing the infectivity of ZIP and ZAP DNA upon transfection into NIH/3T3 cells and for other assistance.

This work was supported by Public Health Service grants CA-26717 and CA-14051 (core grant to S. E. Luria) from the National Cancer Institute. C.S. is a postdoctoral fellow of the National Cancer Institute. D.B. is an American Cancer Society Research Professor.

#### LITERATURE CITED

- Benton, W. D., and R. W. Davis. 1977. Screening  $\lambda$ gt recombinant clones by hybridization to single plaques *in situ*. *Science* 196:180-182.
- Blattner, F. R., A. E. Blechl, K. Denniston-Thompson, H. E. Faber, J. E. Richards, J. L. Slighton, P. W. Tucker, and O. Smithies. 1978. Cloning human



**FIG. 5. Model to explain the existence of the inversion-containing clone 3S (Fig. 2). This model presumes the existence of a head-to-tail dimer circle of M-MuLV DNA reverse transcripts. It is not necessarily a requirement to have two LTRs adjacent to one another in this preintegrative structure (27). A, Auto-integration occurs at the site designated by a dotted line. B, Circular dimer with an inverted segment results. C, Upon *Hind* III digestion and cloning into Charon 21A phage DNA, an insert with the structure of clone 3S is generated.**

- fetal gamma globulin and mouse alpha type globin DNA; preparation and screening of shotgun collections. *Science* **202**:1279-1284.
3. Cohen, J. C., P. R. Shank, V. L. Morris, R. Cardiff, and H. E. Varmus. 1979. Integration of the DNA of mouse mammary tumor virus-infected normal and neoplastic tissue of the mouse. *Cell* **16**:333-345.
  4. Dhar, R., W. L. McClements, L. W. Enquist, and G. F. Vande Woude. 1980. Terminal repeated sequence (TRS) of integrated Moloney sarcoma provirus: nucleotide sequence of TRS and its host and viral junctions. *Proc. Natl. Acad. Sci. U.S.A.* **77**:3937-3941.
  5. Dunsmuir, P., W. Brorein, M. Simon, and G. Rubin. 1980. Insertion of the *Drosophila* transposable element copia generates a 5 base pair duplication. *Cell* **21**:575-579.
  6. Enquist, L., D. Tiemeier, P. Leder, R. Weisberg, and N. Sternberg. 1976. Safer derivatives of bacteriophage  $\lambda$ -C for use in cloning of recombinant DNA molecules. *Nature (London)* **259**:596-598.
  7. Farabaugh, P. J., and G. R. Fink. 1980. Insertion of the eukaryotic transposable element Ty1 creates a 5bp duplication. *Nature (London)* **286**:352-356.
  8. Gafner, J., and P. Philipsen. 1980. Common features of transposition. A yeast transposon also generated duplication of the target sequence. *Nature (London)* **286**:414-418.
  9. Gilboa, E., S. Goff, A. Shields, F. Yoshimura, S. Mitra, and D. Baltimore. 1979. *In vitro* synthesis of a 9 kbp terminally redundant DNA carrying the infectivity of Moloney murine leukemia virus. *Cell* **16**:863-874.
  10. Hirt, B. 1967. Selective extraction of polyoma DNA from infected mouse cell cultures. *J. Mol. Biol.* **26**:365-371.
  11. Hughes, S. H., P. R. Shank, D. H. Spector, H. Kung, J. M. Bishop, H. E. Varmus, P. K. Vogt, and M. L. Breitman. 1978. Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. *Cell* **15**:1397-1410.
  12. Jolicoeur, P., and E. Rassart. 1980. Effect of Fv-1 gene product on synthesis of linear and supercoiled viral DNA infected with murine leukemia virus. *J. Virol.* **33**: 183-195.
  13. Kleckner, N. 1977. Translocatable elements in prokaryotes. *Cell* **11**:11-23.
  14. Kung, H.-J., P. R. Shank, J. M. Bishop, and H. E. Varmus. 1980. Identification and characterization of dimeric and trimeric circular forms of avian sarcoma virus-specific DNA. *Virology* **103**:425-433.
  15. Majors, J. E., and H. E. Varmus. 1981. Nucleotide sequences at host-proviral junctions for mouse mammary tumour virus. *Nature (London)* **289**:253-258.
  16. Maniatis, T., R. C. Hardison, E. Lacy, J. Lauer, C. O'Connell, D. Quon, G. K. Sim, and A. Efstratiadis. 1978. The isolation of structural genes from libraries of eucaryotic DNA. *Cell* **15**:687-701.
  17. Maxam, A., and W. Gilbert. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* **74**:560-564.
  18. Queen, C. L., and L. J. Korn. 1980. Computer analysis of nucleic acids and proteins. *Methods Enzymol.* **65**: 595-609.
  19. Sabran, T. L., T. E. Hsu, C. Yeater, A. Kaji, W. S. Mason, and J. M. Taylor. 1979. Analysis of integrated avian RNA tumor virus DNA in transformed chicken, duck and quail fibroblasts. *J. Virol.* **29**:170-178.
  20. Shank, P. R., S. H. Hughes, H. Kung, J. E. Majors, N. Quintrell, R. V. Guntaka, J. M. Bishop, and H. E. Varmus. 1978. Mapping unintegrated avian sarcoma virus DNA: termini of linear DNA bears 300 nucleotides present once or twice in two species of circular DNA. *Cell* **15**:1383-1395.
  21. Shank, P. R., and H. E. Varmus. 1978. Virus-specific DNA in the cytoplasm of avian sarcoma virus-infected cells is a precursor to covalently closed circular viral DNA in the nucleus. *J. Virol.* **25**:104-114.
  22. Shapiro, J. A. 1979. Molecular model for the transposition and replication of bacteriophage mu and other transposable elements. *Proc. Natl. Acad. Sci. U.S.A.* **76**: 1933-1932.
  23. Shields, A., O. N. Witte, E. Rothenberg, and D. Baltimore. 1978. High frequency of aberrant expression of Moloney murine leukemia virus in clonal infections. *Cell* **14**:601-609.
  24. Shimotohno, K., S. Mizutani, and H. M. Temin. 1980. Sequence of retrovirus provirus resembles that of bacterial transposable elements. *Nature (London)* **285**: 550-554.
  25. Shimotohno, K., and H. M. Temin. 1980. No apparent nucleotide sequence specificity in cellular DNA juxtaposed to retrovirus proviruses. *Proc. Natl. Acad. Sci. U.S.A.* **77**:7357-7361.
  26. Shoemaker, C., S. Goff, E. Gilboa, M. Paskind, S. W. Mitra, and D. Baltimore. 1980. Structure of a cloned circular Moloney murine leukemia virus DNA molecule containing an inverted segment: implications for retrovirus integration. *Proc. Natl. Acad. Sci. U.S.A.* **77**:3932-3936.
  27. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503-517.
  28. Steffen, D., and R. A. Weinberg. 1978. The integrated genome of murine leukemia virus. *Cell* **15**:1003-1010.
  29. Sternberg, N., D. Tiemeier, and L. Enquist. 1977. *In vitro* packaging of a  $\lambda$  dam vector containing Eco RI DNA fragments of *Escherichia coli* and phage P1. *Gene* **1**:255-280.
  30. Swanstrom, R., W. J. DeLorde, J. M. Bishop, and H. E. Varmus. 1981. Nucleotide sequence of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements. *Proc. Natl. Acad. Sci. U.S.A.* **78**: 124-128.
  31. Vogelstein, B., and D. Gillespie. 1979. Preparative and analytical purification of DNA from agarose. *Proc. Natl. Acad. Sci. U.S.A.* **76**:615-619.
  32. Yang, W. K., J. O. Kiggans, D. Yang, C. Ou, R. W. Tennant, A. Brown, and R. H. Bassin. 1980. Synthesis and circularization of N- and B-tropic retroviral DNA in Fv-1 permissive and restrictive mouse cells. *Proc. Natl. Acad. Sci. U.S.A.* **77**:2994-2998.
  33. Yoshimura, F. K., and R. A. Weinberg. 1979. Restriction endonuclease cleavage of linear and closed circular murine leukemia viral DNA's: discovery of a smaller circular form. *Cell* **16**:323-332.