Long terminal repeat (LTR)-derived recombination of retroviral DNA: sequence analyses of an aberrant clone of baboon endogenous virus DNA which carries an inversion from the LTR to the gag region

Taka-aki Tamura and Toshiya Takano

Department of Microbiology, Keio University School of Medicine, Tokyo, Japan

Received 28 June 1982; Accepted 27 July 1982

ABSTRACT
Nucleotidase sequences of a cloned proviral DNA of baboon endogenous virus M7 were analyzed, which carried an internal inversion. The inversion of 2.2 kilobase pairs was occurred between the junction of two tandem LTRs and a site locating in the p30 region of the gag gene. The ATAA sequence was a target for recombination generating the inversion, which was duplicated at both ends of the inverted segment. AAA and CA were lost at the 5'- and 3'-ends of the LTRs by the inversion, respectively. On both sides of the target sequence, long AG-rich stretches were detected, which may specify the site of recombination together with the target sequence. The characteristic base changes in the inversion are concluded to result from an illegitimate recombination associated with LTRs, as well as in case of provirus integration into the host cell DNA. We propose and discuss models to explain the processes of recombination to generate both inversion and integration.

INTRODUCTION
During the life cycle of retroviruses, the viral RNA is transcribed into DNA by reverse transcriptase, and three forms of the viral DNAs (proviruses) are detected in the infected cell (1). One of the DNA transcripts is a linear molecule flanked with directly repeated sequences, LTRs, at each end, which consist of the 3'- and 5'-end sequences of the viral RNA genome (2, 3). The other provirus DNAs have circular configuration containing either one or two copies of the LTR. The integrated form of the provirus is detected in the host cell DNA, which is colinear with the unintegrated linear provirus (4). The integrative recombination with the host DNA is suggested to occur at the LTR termini (3, 4, 5). The immediate precursor of the integration has not been identified. The biological functions of the LTR are not clearly understood, but its structure warrants several speculations: The LTR contains control elements for promotion and termination of transcription, and may mediate or facilitate integration of the viral DNA to the host chromosome (6, 7, 8, 9). Though the provirus integration is one of the crucial steps for establishment of retroviral infection, the mechanism of the integration has not been understood.
at molecular levels yet. However, the following changes have been generally observed on the nucleotide sequences at a site of integration: Two base pairs (bps) of the LTR termini are usually lost and several bps become reiterated on the chromosome DNA (10). The structure and functions of retroviral proviruses resemble to those of transposable elements in bacteria (11), the copia and related sequences in Drosophila (12) and Ty1 elements in yeast (13).

We have recently cloned the circular provirus DNAs of baboon endogenous virus M7 (14) and have analyzed the nucleotide sequences of the M7 LTR (6). Among the M7 DNA clones isolated, a particular one, λBEV-27, carries an internal inversion which ends at the termini of the LTRs. Shoemaker et al. (15, 16) and Van Beveren et al. (17) reported the similar inverted clones of the proviral DNAs of Moloney murine leukemia and sarcoma viruses, respectively. From the DNA sequences of these clones, the inversions were suggested to be generated by recombination events on the usual, non-inverted provirus DNAs. In this report, we analyzed the nucleotide sequences of the regions where the inversion occurred, and those of the non-inverted clone of the M7 DNA. The inversion of λBEV-27 was concluded to occur just at the junction of the two tandem LTRs, and to extend to the p30 region of the gag gene. The sequences around the target site in the p30 region, which are rich in purine nucleotides, were homologous to those of Moloney leukemia virus. From these results, we discuss the function of retroviral LTR and the mechanisms of the recombination which generates inversion and integration of the proviruses.

MATERIALS AND METHODS

Subcloning of M7 DNA Fragments. Two recombinant phages, λBEV-11 and λBEV-27 (14), were used. λBEV-11 carries the non-inverted M7 DNA of a full genome length, 8.6 kilobase pairs (kb), while λBEV-27 carries the inverted provirus of the same, full genome length. The restriction maps of the clones are shown in Fig. 1. The Sal I-, Hind III-fragment of about 1.6 kb, which covered the gag gene region, was subcloned from the insert of λBEV-11 into a plasmid pBR322. A Bam HI fragment of 2.0 kb with one LTR, and a Bam HI-, Eco RI-fragment of 2.7 kb containing the inverted segment, were also subcloned from λBEV-27 to pBR322.

Sequence Analyses. The chemical modification method of DNA sequencing described by Maxam and Gilbert (18). The detailed conditions for the chemical modification and electrophoresis were described before (6).

RESULTS

Nucleotide Sequences around the Inversion Endpoints of λBEV-27. A recombinant
Fig. 1. Restriction maps of the non-inverted and inverted M7 proviruses in λBEV-11 and λBEV-27 and the construction of the subclones. Sites for Eco RI (▲), Bam HI (●), Sal I (○), Hind III (▲), Pst I (□) and Xho I (■) are indicated by vertical lines. The LTRs are shown as rectangles. The 1.6, 2.0 and 2.7 kb fragments of λBEV-11 and λBEV-27 were subcloned as shown in the figure and the recombinant pBR322 were designated as pBE-SH, -B2 and -BE, respectively. The regions A, B, C and D of the cloned DNAs, shown by blank arrows, were sequenced.

Phage, λBEV-27, has been isolated in our earlier study (14), which carries a novel inverted structure of the M7 DNA insert. The physical map of λBEV-27 is shown in Fig. 1. The segment from the junction of the two tandem LTRs to the gag gene region was internally inverted in this clone. By transfection with calcium phosphate precipitation method (19), the insert M7 DNA of this phage was not able to direct any progeny virus production, even after the transfected cells, A204 (20), were passaged for 60 days (data not shown). Whereas, the progeny virus production was detected in the cultures transfected with the insert M7 DNAs of the non-inverted clones carrying one or two LTRs, on about 60 days after transfection.

From the sequence of the left end of the inverted segment of λBEV-27 (Fig. 2B), the endpoint of the inversion was detected at the edge of the LTR sequence, about 100 bps upstream from the right Bam HI site (see Fig. 1). Comparing
Fig. 2. Nucleotide sequences of the endpoints of the inverted segment in λBEV-27 and those of the corresponding sites in λBEV-11. The locations of the sequences, A, B, C and D are shown in Fig. 1 by the blank arrows, A, B, C and D, respectively. The sequences are presented in the same directions as those of the blank arrows of Fig. 1. The stretches enclosed by thin lines are the LTR sequences detected in each region. Heavily underlined sequence is the target site for recombination in the inversion (D), which duplicated at each end of the inverted segment (B and C).

with the sequence of LTR found in the non-inverted clone, λBEV-11, (Fig. 2A), one base pair (bp), G, was deleted at the recombination site. At the right end of the inverted segment (Fig. 2C), another LTR terminus appeared and it again lacks three bps, TTT. We concluded that the recombination in the inversion of λBEV-27 should occur just at the ends of the LTR sequences and the deleted bps were lost during the recombination process. On the other hand, four bp sequences, ATAA, were commonly found at both of the recombination sites (Fig. 2B and C), which seemed to be the target of the recombination.

Nucleotide Sequences adjacent to the Target Site of the Recombination. In the restriction maps of the M7 inserts from λBEV-11 and -27, the Bam HI site of λBEV-27, locating just at the right-hand side of the LTR, corresponded to that of the non-inverted clone which located 2.2 kb downstream from the tandem LTRs (Fig. 1). The region around the Bam HI site of λBEV-11 is in the gag gene of M7. A recombinant plasmid pBR322, pBE-SH, which carried a 1.6 kb Sal I-, Hind III-fragment of λBEV-11 (Fig. 1), was constructed and used for sequencing (Fig. 3). The sequence was characteristically rich in purine nucleotides and more than 70% of the bases were A or G. This sequence was split into two parts in the inverted clone by recombination, which appeared downstream from the target in Fig. 2B and C. Interestingly, the target site, ATAA, was detected once in the original sequence of the non-inverted clone (Fig. 2D and 3), while the ATAA sequence was duplicated at the recombination sites of the inverted clone (Fig. 2B and C).

Open Reading Frames around the Target Site. As described above, the inversion seemed to occur by recombination at the target site in the gag gene and the biological activity of the inverted M7 DNA was inert to direct the progeny.
Fig. 3A. Nucleotide sequence of the region around the target site in λBEV-11. The location of the sequence in the insert of λBEV-11 is shown in Fig. 1 by the blank arrow D. A portion of the sequence is also shown in Fig. 2D. The nucleotide sequence is presented by one of the two strands which has the same polarity as the viral genome. The nucleotides are numbered from the 5'-end of the Bam HI site located in this region (see Fig. 1). Nucleotide stretches containing more than 5 purine nucleotides are underlined. The amino acids shown under the nucleotide sequence are deduced from the open reading frame 2 of the nucleotide sequence. B. Lists of the termination codons detected in the nucleotide sequence shown in A, according to the three reading frames. The numbers indicate the positions of the first base of each termination codon in the nucleotide sequence shown in A. AM, amber (TAG), OP, opal (TGA), OC, ochre (TAA) codons are shown.

production. We searched for possible open reading frames in the region around the target site of λBEV-11. Three termination codons, amber (TAG), ochre (TAA) and opal (TGA), were detected in the sequence, as summarized in Fig. 3B. In the reading frames 1 and 3, too many termination codons are located in the region, while the frame 2 does not give any terminations. The amino acid sequence was deduced from the nucleotide sequence assigned by the frame 2 (Fig. 3A). On the other hand, four kinds of the gag proteins of M7, NH2-p12-p15-p30
-p10-COOH, have been identified and the amino acid sequence was partially determined by Copeland et al. Our data of the amino acid sequence shown in Fig. 3A are well coincident with their results in the region around the junction of p30 and p10, although mismatches of several amino acids are observed. From these results, the junction of p30 and p10 seems to be located between the leucine at 52nd and the alanine residues at 53rd positions. From the amino acid analyses on gag proteins of various retroviruses reported so far (21, 22, 23, 24), the first amino acid of the fourth gag gene protein is generally an alanine residue. From our sequence data (not shown), the nucleotide stretch coding p30 seemed to start at the 127th base downstream from the unique Sal I site of the M7 DNA. The conclusion that the last amino acid of p30 is the leucine residue at the 52nd position is not inconsistent with these findings. Therefore, the target sequence for recombination in the inversion of λBEV-27 should locate 66 bps upstream from the end of the p30 region (Fig. 3A). These data well explain the deficiency of infectivity by transfection with the M7 insert of λBEV-27.

DISCUSSION

Inversion in Provirus DNA of Retroviruses. An aberrant form of the proviral DNA of M7 has been cloned in λBEV-27. The size of the proviral DNA is 8.6 kb as well as that of the non-inverted provirus of M7, while this clone carries a 2.2 kb internal inversion extended from the junction of the two tandem LTRs to the p30 region of the gag gene. Shoemaker et al. (15, 16) and Van Beveren et al. (17) have isolated similar inverted clones of the provirus DNAs in Moloney leukemia and sarcoma viruses. They reported that the inversions always occurred at the junction of the tandem LTRs on the circular proviral DNAs. This type of the proviral inversions likely occurred during synthesis of the viral DNA in the virus-infected cells, but not during growth of the recombinant phages in the bacterial host, because we have not found any inversion nor deletion mutants of λBEV clones during multiple passages on the bacteria. The frequency of detecting these aberrant proviral DNAs are significantly high (16, 17, 25). In our case, an inverted clone was detected among 17 recombinant phages carrying the provirus of M7.

The Inversion of λBEV-27 was Originated by Illegitimate Recombination. From the nucleotide sequence data, the target sequence for recombination was deduced to be ATAA in the inversion of λBEV-27. The sequences around the target site were not homologous with those of each counterpart of the recombination. Therefore, the inversion was resulted from illegitimate recombinations. In
the inversion, the following characteristic features are observed; first, one end of the inverted segment is located at the junction of the LTRs. Second, the target site of several bps are reiterated. Third, two or three bps are lost from the LTR termini. Similar base changes have been reported in provirus integration of various retroviruses (10). Therefore, the processes of recombination to generate the inversion in λBEV-27 must be analogous to those in the proviral integration into the host DNA, i.e., the inverted structure had resulted from a circular viral DNA molecule integrating within itself. Both inversion and integration of retroviral proviruses should be catalyzed by recombination functions stipulated on an eukaryotic transposable element, LTR.

Precursor Provirus DNAs and Processes of Inversion. Although three kinds of provirus DNAs are known to be present in the newly infected cells, immediate precursor for integration has not been determined. In the inversion of the M7 provirus, either one of the three viral DNAs can be the candidate for the precursor. One of the simplest assumptions is that the large circular form of the provirus carrying two tandem LTRs might be converted to an inverted DNA molecule by intramolecular (Fig. 4A) or intermolecular recombinations (Fig. 4C). The amounts of the circular proviral DNAs are limited in the newly infected cells, hence the inversion by the intermolecular recombination seems to be less likely. Saedler et al. (26) extended Shapiro's model (27) and have proposed models for bacterial transposons which explain the generation of inversion and deletion. According to Saedler's model, the small circular provirus DNA can be a precursor (Fig. 4D). In this case, however, DNA replication has to occur through the whole stretch of the LTR. In the intramolecular recombination model described above, DNA synthesis is required only on the sequence of a few bases at the LTR junction. On the other hand, a dimer molecule of the large circular provirus is suggested as a precursor of inversion (16). This assumption may also explain the generation of inversion in λBEV-27, but seems unlikely because the inversions of proviruses may not be generated at a high frequency from a trace amount of the dimer molecules in the population of the provirus DNAs. We have not yet detected any dimer molecules of the M7 provirus in the infected cells. The another alternative of the precursor might be the linear form of the provirus. The data presented here cannot exclude either one of these hypotheses.

Recombination Processes to Generate Inversion and Integration of Provirus DNA. We propose a model to explain the recombination processes generating both inversion and integration of the M7 provirus (Fig. 5), which is essentially similar to that proposed by Shapiro (27). In the model, we assume the large
Possible precursors and processes of the inversion in \( \lambda \)BEV-27 are schematized. The LTR is shown by open or solid rectangles and the recombination indicated by \( X \) occur between the termini of the LTR and the target sites shown by arrows and black circles. From the large circular form of the provirus DNA, an inversion (A) or a deletion (B) can be generated by intramolecular recombination. In C, an intermolecular recombination occurs on two molecules of the large circular provirus. The small circular form of the provirus DNA can be the precursor (D).

circular form of the provirus with two tandem LTRs, as the precursor of the recombination. First, a certain enzyme makes staggered cuts at the junction of the two LTRs and at the target site in opposite orientations (Fig. 5-1). At the LTR junction, the DNA is nicked specifically at the 5'-ends of ..GT sequences of both strands (the cuts 1 and 2 in Fig. 5-1). The GT is the specific sequence commonly found in the inverted repeats of eukaryotic and prokaryotic transposable elements (10). The other staggered cuts occur four bases apart at each edge of the target sequence. Second, each protruding end of the target sequence, 3'-AATA and TTAT-3', is jointed to the 5'-TG terminus of the recessed ends of the LTR sequences (Fig. 5-2). This ligation of the overhanging strands of the target site results in mismatching of the five bases, 3'-AAACT and 3'-AGT, derived from the junction of the LTRs, which are excised by a nuclease (Fig. 5-3). Finally, the single-stranded gaps are fil-
Fig. 5. A model to explain the recombination processes in inversion and integration of the M7 provirus. The recombination is assumed to occur between the target site and the junction of tandem LTRs of the large circular provirus.

1. formation of staggered cuts on each counterpart of the recombination. The small arrows numbered as 1 to 4 indicate the position where the staggered cuts are made; 2. joining of the protruding 3'-ends of the target sequence to the 5'-TG termini of the LTRs; 3. elimination of the single-stranded 5 bp stretches from the other LTR termini; 4. filling the single-stranded gaps.

led by DNA synthesis with a repair enzyme and the target sequence are reiterated (Fig. 5-4). Similar recombination may be processed in integration of the provirus to a target on the host DNA, as well as those suggested in integration of the proviruses in many other retroviruses (10).

In the recombination process proposed above, the step making staggered cuts at the specific sites in the inverted repeats of the LTRs seems to be unique and essential for the recombination in transposable elements. In bacterial transposons, transposase or resolvase which catalyzes a site specific nicking activity has recently been isolated (28, 29) and the structural gene of the enzyme is coded by the transposons themselves (30, 31). In eukaryotic systems, however, similar enzymes have not been identified yet. One of the retroviral gene products or a certain unidentified protein coded by the virus genome might harbor similar enzyme activities. In this paper, the structure and the specificity of the substrate DNA molecules have been identified so that such a hypothetical enzyme can be isolated by using in vitro assay systems.

Characteristics of the Nucleotide Sequences around the Target Site. Proviruses are usually integrated at random sites on the host chromosome (16, 32).
As for the target sites of several bacterial and eukaryotic transposable elements, the sequences are generally AT-rich (32, 33, 34). In the inversion of ABEV-27, the target sequence was ATAA TATT. The ten base stretch at the target site, CTTATAAAG, can form a hairpin configuration. This sequence is strikingly similar to that of the target site, TTATAA, of a bacterial transposon γ (29). The AT-rich and palindromic structure of these sequences might be essential for the recognition by the enzyme analogous to the transposase. On the other hand, the long sequences detected on both sides of the target site were rich in purine nucleotides (Fig. 3). In one of the inverted provirus DNA of Moloney leukemia virus, the target sequence, TCAG, was located in the p30 region of the gag gene and long AG-rich sequences were also detected around the target site (15). The AG-rich M7 sequences of 60 and 16 bps which locate upstream and downstream of the target site, respectively, are completely coincident with those of Moloney leukemia virus. Therefore, these long AG-rich regions were conserved in these two viruses, which might specify the sites of recombination together with the target sequences.

ACKNOWLEDGMENTS This work was supported by research grants from the Japanese Ministry of Education, Science and Culture, the Naito Foundation, and the Waksman Foundation of Japan, and by the institutional grant.

REFERENCES