**Nucleic Acids Research**

**Target sites for the transposition of rat long interspersed repeated DNA elements (LINEs) are not random**

Anthony V.Furano, Charles C.Somerville*, Philip N.Tsichlis[1] and Ettore D'Ambrosio

Section on Genomic Structure and Function, Laboratory of Biochemical Pharmacology, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892 and [1]Fox Chase Cancer Center, Philadelphia, PA 19111, USA

ABSTRACT

The long interspersed repeated DNA family of rats (LINE or L1Rn family) contains about 40,000 6.7-kilobase (kb) long members (1). LINE members may be currently mobile since their presence or absence causes allelic variation at three single copy loci (2, 3): insulin 1, Moloney leukemia virus integration 2 (Mlvi-2) (4), and immunoglobulin heavy chain (Igh). To characterize target sites for LINE insertion, we compared the DNA sequences of the unoccupied Mlvi-2 target site, its LINE-containing allele, and several other LINE-containing sites. Although not homologous overall, the target sites share three characteristics: First, depending on the site, they are from 68% to 86% (A+T) compared to 58% (A+T) for total rat DNA (5). Depending on the site, a 7- to 15-bp target site sequence becomes duplicated and flanks the inserted LINE member. The second is a version (0 or 1 mismatch) of the hexanucleotide, TACTCA, which is also present in the LINE member, in a highly conserved region located just before the A-rich right end of the LINE member. The third is a stretch of alternating purine/pyrimidine (PQ). The A-rich right ends of different LINE members vary in length and composition, and the sequence of a particularly long one suggests that it contains the A-rich target site from a previous transposition.

INTRODUCTION

Long interspersed repeated DNA families [LINE (6) or L1 families (7)] are one

class of highly repeated DNA families that together account for a third or

more of various mammalian genomes (6, 8). Although the evolutionary origin

and present day function of LINE DNA is not known, this class of sequences,

like other highly repeated interspersed sequences, must have possessed at

some time, or have been unusually susceptible to, one or more means of

amplification and transposition. Related but distinct LINE families are

present in rats (1), mice, and primates (9, 10). Therefore, amplification of

different ancestral LINE families has occurred during the evolution of these

species. Since the rat LINE family is much more homogeneous (1) than those

in mice or primates (7, 9-11), it seems that the rat LINE family has been

amplified, or rectified, quite recently. Other results indicate that LINE

transposition may still be rather active in rats: First, at least three

single copy loci (Mlvi-2, Igh, insulin 1) of Rattus norvegicus are poly-
morphic due to the presence or absence of LINE sequences (2, 3). Second, the
Igh locus of the closely related species, R. rattus, is also polymorphic due
to the insertion of a LINE member. Since the site of insertion in the Igh
gene was different in the two species, these LINE insertions were independent
events [(2) and B. L. Schaffer, unpublished observations].

The factors that influence LINE transposition are largely unknown.
However, transposition of LINE members to new genomic sites could have impor-
tant biological consequences. The newly resident LINE member could not only
affect the activity of genes in the contiguous DNA but also promote its rear-
rangement by recombination with other LINE members in the genome [e.g., see
(12)]. Therefore, LINE mobility could be a major cause of genetic novelty
and diversity in an animal population.

The identification of LINE-related polymorphisms in rat permits compari-
son of the DNA sequence of a target site before and after its invasion by a
LINE member. Therefore, the characteristics of a LINE insertion site and the
effects of the LINE insertion on it can be directly determined rather than
inferred. Here we compared the "empty" Mlvi-2 target site with its LINE-
containing allele and found that the LINE which is inserted at Mlvi-2 is
flanked by a duplicated, 14-bp target site sequence. This is the only major
change in the target site ascribable to the insertion. Comparison of the
LINE-containing Mlvi-2 allele with other LINE-containing sites revealed that
the target sites, while not homologous overall, share three characteristics.
Since each characteristic is also present in the empty Mlvi-2 site, none is
either a consequence of, or occurred subsequent to LINE insertion. We
estimate that the number of these target sites is such that their juxta-
position to LINE members could not have occurred by chance.


MATERIALS AND METHODS
DNA Sequence Determination
The appropriate DNA fragments (see Fig. 1) of the previously reported (2)
clones of the empty and LINE-containing Mlvi-2 alleles were subcloned into
pUC9 (13). The inserts were released by digestion with the appropriate
restriction endonuclease(s) and purified by agarose gel electrophoresis.
Both DNA strands of the indicated regions (see legend to Fig. 2) were se-
quenced using a modification (14) of the chemical degradation procedure (15)
or by the chain termination procedure (16) using M13 vectors (17). When
necessary, ExoIII deletion clones were prepared as described by Henikoff
(18).

## Statistical Analysis

The probability (denoted as $P_1$) that a TACTCA or any of its 18 singly sub-
stituted variants occurs by chance in a segment of DNA was calculated as
follows: First, the chance probabilities of finding either TACTCA, or each
variant, at a single position in a given DNA segment were calculated from its
base composition. Then, the sum of these probabilities was subtracted from
1, and the remainder was raised to the power of (length of DNA segment - 6).
This value is the probability of not finding any of the hexamers at any
position in the DNA segment. Subtraction of this value from 1 gives the
probability, $P_t$, that TACTCA or any of its variants occurs 1 or more times in
the DNA segment. We note that $P_t = P_1 + P_2 + P_3 \ldots P_i$, where $P_i$ is the
probability that TACTCA or any of its variants occurs in the DNA segment i
times. If we assume that $P_i \cong P_1^i$, then $P_t = \sum_{i=1}^{\alpha} P_1^i$. Since $\sum_{i=0}^{\infty} P_1^i = 1/(1 - P_1)$
(19), it follows that $P_1 \cong P_t/(1 + P_t)$.

## RESULTS

Fig. 1 shows a diagram of a typical rat LINE family member (1) and of the
LINE member inserted at the Mlvi-2 locus (2). The Mlvi-2 LINE is colinear
with the typical member but is 31 bp shorter at the left end (see below).
Fig. 1 also shows the empty target site cloned from the allele without the
inserted LINE. Fig. 2A shows the DNA sequence of the empty target site.
Three sequence elements are indicated: The first is a 14-bp sequence (boxed)
that contains a stretch of A's. This sequence becomes duplicated upon LINE
insertion (see below). The second is the hexanucleotide, TACTCA (underlined
and italicized). The third is a perfect stretch of 26 TG pairs (underlined),
also referred to below as a PQ stretch.

Fig. 2B shows the left junction between non-LINE DNA and the Mlvi-2
LINE. The lefthand member of the duplicated target site sequence is boxed.
Just to the right of this sequence is the left end of the Mlvi-2 LINE;
dashes have been inserted here to compensate for the 31 bp that are missing
from the left end of the Mlvi-2 LINE. Fig. 2C shows the right LINE/non-LINE
DNA junction. The A-rich right end of the Mlvi-2 LINE is contiguous with the
righthand member of the duplicated target site sequence (boxed), and just to
the right of it is the hexanucleotide, TACTCA, and the TG stretch identified
in the empty target site (Fig. 2A). The TG stretch in the LINE-containing
site is shorter than that in the empty site. This length difference is not
unexpected since "simple" sequences, such as a TG stretch, can either expand
or contract by unequal crossing over or by "slippage" during DNA replication
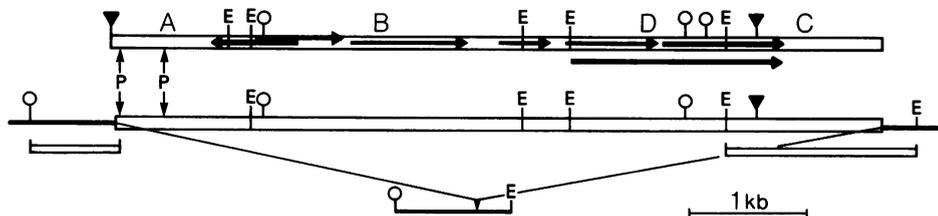
FIG. 1. Rat LINE family members. The top diagram shows a typical rat LINE family member (1). The heavy filled arrows within the open bar indicate open reading frames found in a sequenced randomly selected full length member of the family (called LINE 3), and the heavy filled arrow on the right below the open bar indicates an ORF in another randomly selected LINE member (called LINE 4) (1). These righthand ORFs correspond to the ORFs in the mouse LINE sequence (32). Two BamHI sites (▼) delimit a 5.5-kb fragment present in at least half of the genomic copies of this family. The highly conserved EcoRI (E), HindIII (⌀), and PstI (P) sites are also shown, and the letters A, B, D, and C refer to LINE segments that correspond to the main EcoRI fragments present in most family members (1). The middle diagram shows the relevant restriction enzyme sites from a previously published map (2) of the LINE member inserted at the Mlvi-2 locus, and the open bars underneath it show the cloned fragments containing the left and right junctions. The bottom diagram shows the cloned fragment of the empty target site.

(20). Fig. 2C also shows that the target site hexanucleotide, TACTCA, is also present in the inserted LINE sequence, four nucleotides from the beginning of its A-rich right end.

Fig. 3A shows a comparison of the right Mlvi-2 LINE/non-LINE junction with the corresponding junctions of two randomly selected rat LINE members (LINEs 4 and 3) that we previously sequenced (1). LINE 3 is flanked by a 14-bp A-rich perfect direct repeat (boxed). Beyond the end of LINE 4 is a 1 mismatch version of the hexanucleotide, TACTCA, and 66 bp beyond this, a perfect stretch of 16 GT pairs. Beyond the end of LINE 3 there are two single mismatch versions of the LINE hexamer: the first, TGCTCA, which is italicized and underlined, is followed after one base by the second, TACTTA. An imperfect stretch of 21 AT pairs occurs 256 bp beyond the second hexamer. These results not only show that versions of the sequence elements identified in the empty Mlvi-2 target site are also present in randomly selected LINE-containing sites, but that the orientation of the target site elements both to each other and the inserted LINE member is the same in all three cases.

Fig. 3B shows the only other two sequenced examples of the junction between non-LINE DNA and the intact right end of a rat LINE member. In contrast to the LINE members shown in Fig. 3A, LINEs I1 and RC12 consist of only the D and C segments (I1) [(3) and M. Soares et al., personal communication]

## A target site

```
CATATGATAGTTTTTATTTTCCATCTTCTGAGGAACAATTATACTGATTTC    50
CATAACAATGAAAAGTTGATTTGAAGTCCCATCAATGGTTAATATGGGTT    100
TCTTCACCACATAATCATCATTGTTTGTCTTCTTTTAAAAAAATGATTTG    150
TGTATTTATATGTAGTACTCATCTGTGTGTGTGTGTGTGTGTGTGTGTGT    200
GTGTGTGTGTGTGTGTGTGTGTGTACTGTTTGTATGCCTGGTGCCTGC    250
TGAGGTCAGAAGAGGATAATGAATACCCTAGAGGTATAGTTACAGACAAT    300
TATAAGCTGCTGCATGGATTCTTAGAATCCGGTGCAGATCCTCTGCAAGA    350
GCAGCAAGTGCTCTATCATGCTGAGCCCTCTCTCCTGGCCCTAGTTTACC    400
TTCTCGACGATAATCA                                      416
```

## B left junction

```
•••••••••••••••••••••••••••••••••••••••••••••••••••
CATATGATAGTTTTTATTTTCCATCTTCTGAGGAACAATTATACTGATTTC    50
•••••••••••••••••••••••••••••••••••••••••••••••••••
CATAACAATGAAAAGTTGATTTGAAGTCCCATCAATGGTTAATATGGGTT    100
•••••••••••••••••••••••••••••••••••••••••••••••••••
TCTTCACCACATAATCATCATTGTTTGTCTTCTTTTAAAAAAATGATTTG    150
----------------------------------CTCCCAGACCCGGTGAGAG    200
CCCGAGGATCCCGGCCCCGCAGCAGCTCTCTG.C.........T.......
AGAGACCCAACCGCCTGGTCAGGTGGGCACTCCTGAGGCTGCAG          244
```

## C right junction

```
GAATTCACAGCTGAGGAATGCCGAATGGCTGAGAAACACCTAAAGAAATG    50
............................................
TTCAACATCTTTAGTCATAAGGGAAATGCAAATCAAAACAACCCTGAGAT    100
............................................
TTCACCTCACACCAGTGAGAATGGCTAAGATCAAAAACTCAGGTGACAGC    150
...............GT......................
AGATGCTGGCGAGGATGTGGAGAAAGAGGAACACTCCTCCATTGTTGGTG    200
............................................
GGATTGCAGACTGGTAAAACCATTCTGGAAATCAGTCTGGAGGTTCCTCA    250
............................................
GAAAATTGGACATTGAACTGCCTGAGGATCCAGCTATACCTCTCTTGGGC    300
............................................
ATATACCCAAAACATGCCTCAACATATAAAGAGACACGTGCTCCACTAT    350
..................G.........................
GTTCATCGCAGCCTTATTTATAATAGCCAGAAGCTGGAAAGAACCCAGAT    400
.....T..C..A...............A.................
GCCCTTCAACAGAGGAATGGATACAGAAAATGTGGTACATCTCACACAATG    450
............................................
```

```
GAATATTACTCAACTATCAAAAACAACGAGTTTATGAAATTCGTAGGCAA    500
............G.............................
ATGGTTGGAACTGGAAAATATCATCCTGAGTGAGCTAACCCAATCACAGA    550
...............................C.........
AAGACATACATGGTATGCACTCATTGATAAGTGGCTATTAGCCA-AATGC    600
                                   ORF STOP
                                    [.].C.....
TTGAATTACCCTAGATCCCTAGAACAAACGAAACTCAAGACGGATGATCA    650
............................................
AAATGTGAATGCTTCACTCCTTCTTTAAATGAGGAAAAAGAATAC-CTTG    700
............................................C.....
GCAGGGAAGGGAGAGGCAAAGATTAAAACAGAGACTGAAGGAACACCCAT    750
..T.........................................
TCAGAGCCTG-CCCACATGTGGCCCATACATATACAGCCCACCCAATTAGA    800
.........C.....G...........................G..
CAAGATGGATGAAGCAAAGAAGTGCAGACCGACA-GAGCCGGATGTAGAT    850
..............................T...G.........
CGCTCCTGAGAGACACAGCCAGAATACAGCAAATACAGAGGCGAATGCCA    900
............................................
GCAGCAAACCACTGAACTGAGAATAGGTCCCCTATTGAAGGAATCAGAGA    950
..................................CG.........
AAGAACTGAAAGAGCTTGAAGGGGCTCGAGACCCCAAAAGTACAACAATG    1000
.........G.....................A............
CCAAGCAACCAGAGCTTCCAGGGACTAAGCCACTACCTAAAGACTATACA    1050
T...........................................
TGGACTGACCCTGGACTCTGACCCCATAGGTAGCAATGAATATCCTAGTA    1100
............................................
AGAGCACCAGTGGAAGGGGAAGCCCTGGGTCCTGCTAAGACTGAACCCCC    1150
..........................................G...
AGTGAACTAGACTATGGGGGGAGGGCGGCAATGGGGGGAGGGTTGGGAGG    1200
............................................
GGAACA-CCATAAGAAAGGGGAGGGGGGAGGGGGATGTTTGCCCGGAAAC    1250
..G...N.......G...............................A.........
-GAAAGGGAATAAACACTCGAAATGTATATAAGAAATACTCAAGTTAATAA    1300
C.................TA..T--...............
                             T••••••••••••••••••••••
AAAAATAAATAAAAAAAAAATGATTTGCGTATTTATATGTAGTACTCATC    1350
TGTGTGTGTGTGTGTGTGTGTGTGTG•••••••••••••••••••••••••
----------------------------TGTGTGTGTGTGTGTGTGTG    1400
••••••••••••••••••••••••••••••••••••••••••••••••
TGTACTGTTTGTATGCCTGGTGCCTGCTGAGGTCAGAAGAGGATAATGAA    1450
••••••••••••••••••••••••••••••••••••••••••••••••
TACCCTAGAGGTATAGTTACAGACAATTATAAGCTGCTGCATGGATTCTT    1500
A•••••C•••••••••••••••••••••••••••••••••••••••••••
GGAATCGGGTGCAGATCCTCTGCAAGAGCAGCAAGTGCTCTATCATGCTG    1550
••••••••••••••••••••••••••••••••••••••••••••••••
AGCCCTCTCTCCTGGCCCTAGTTTACCTTCTCGACGATAATCATTTTCCT    1600
ATAATGAGACAGAATTC                                    1617
```

FIG. 2. DNA sequence of the empty Mlvi-2 target site and the left and right LINE/non-LINE junctions. Panel A. The sequence begins at an NdeI site, which is in the middle of the cloned target site fragment, and extends to 24 bp from the EcoRI site (see Fig. 1). Panel B. The sequence of the left non-LINE/LINE DNA junction extends from the NdeI site of the target site (see panel A) to the PstI site which is in the LINE element (see Fig. 1). The sequences of the non-LINE flanking DNA and the corresponding part of the empty target site are identical as indicated by the "*"'s above the sequence. The left end of the LINE 3 sequence (1) is shown below the Mlvi-2 LINE sequence by a "●" where the sequences are identical or a letter where they are not. Panel C. The sequence of the entire EcoRI fragment which contains the right Mlvi-2 LINE/non-LINE junction is shown (see Fig. 1). This includes the entire C segment of the Mlvi-2 LINE (cf., upper diagram of Fig. 1), which contains an ORF that stops at the same position as do the corresponding ORFs in other LINEs [(1, 21), M. Soares et al., personal communication]. The right end of the LINE 3 sequence is shown below the Mlvi-2 sequence; the A-rich right end of LINE 3 is not included. Comparison of the non-LINE DNA with the empty target site is shown above the sequence by an "*" where the sequences are the same or a letter where they are different.

or the C segment only (RC12) (21) (compare to Fig. 1, top diagram). In each case, the duplicated sequence that flanks these members contains a stretch of A residues, and a version of TACTCA is present in the non-LINE DNA beyond the

A

```
LINE                10        20        30        40
                     |         |         |         |
Mlvi2  TACTCAAGTT AATAAAAAAAATAAATAAA AAAAAAATGATTTGCGTATT

4      TACTCAAGTT AATAAAAATAAAATAAAA TAAAATAAAAAAGGAGCATT

3      TACTCAAGTT AATTAAAAATAATAATAA CAAAATGGAAACATGAAAA
                   TAATAATAAAAAGAATTA
                   CTCATAAAATAAAGACAA
                   GATGATGGATAAAATTAA
                   AAAAACAAAAAAAACAAA
                          AAAAAAC
```

```
           50        60
            |         |
Mlvi2  TATATGTAGTACTCATCT-(GT)₁₂

4      CAACTCATAAATAATTTT-(55bp)-(GT)₁₆

3      TGCTCACTACTTACAATAG-(250bp)-(AT)₂T(AT)₅A(AT)₄TATA
                                   (AT)₄CATACT(AT)₄
```

B

LINE

```
I1     TACTCAAGTT TAATAAAAAAATAAAAAA ATAAAATTCTCATGTTTTAT
                  ATAAATAAATAA

RC12   TACTCAAGTT AATAAAAAAAAT        GTTTAAAAAAAAAAAAAAAA
```

```
I1     TTTAATTGATTTATTTACTGA-(164bp)
```

```
            ↓
RC12   AAGAAGGGAGAAAGGACA
       GACCTAATAGCTTATCTT
       AAAAAGGCTACTAATGAG
       TAATTCTACTGCCTTATT
       TATTACAAAACAAATGTC
              TCATGGCTTTTA-ATGTATACCATAAT-(235bp)-
                           TATACTTAGATTTACATAAC
                           TGGGTATATGAATATG
```

FIG. 3. Right LINE/non-LINE junctions of various LINE members. Part A shows the DNA sequences of the right end of LINE Mlvi-2, LINE 4 and LINE 3, and the non-LINE DNA just beyond [(1) and this work], and part B shows the corresponding sequences of LINE I1 (M. Soares et al., submitted) and LINE RC12 (21). The target site duplications, which are boxed, are aligned. The target site duplication of the LINE 4 member could not be determined, since this member was fragmented as a result of the initial cloning. In the empty insulin 1 target site, the DNA sequence corresponding to the one boxed in Fig. 3B is ATTAAATTCTCGTGT, and the lefthand member of the duplicated target site is ATTAAATTATCGTGT [(3), M. Lakshmikumaran, unpublished observations]. Imperfect stretches of PQ beyond the right end of LINE RC12 are underlined. The arrow marks the beginning of the cytochrome c pseudogene into which LINE RC12 inserted (21), and the boxed sequence, TAA, is its termination codon.

TABLE 1.  Sequence composition of LINE target sites

| | Left flank [a] | | | Right flank | | | |
|---|---|---|---|---|---|---|---|
| LINE | Fraction (A+T) | Copies of hexamer [b] | $P_1$ | Fraction (A+T) | Copies of hexamer [b] | $P_1$ | $(P_n)$ [c] |
| Mlvi-2 | 0.68 | 1 | 0.19 | 0.81 | 1 | 0.14 | (0.09) |
| 4 | | | | 0.83 | 1 | 0.19 | |
| 3 | 0.63 | 1 | 0.21 | 0.68 | 2 | 0.03 [d] | (0.03) |
| I1 | 0.53 | 0 | 0.24 | 0.86 | 1 | 0.15 | |
| RC12 | 0.43 | 0 | 0.11 | 0.71 | 1 | 0.14 | |

a  These data were derived from the results of Fig. 2A (Mlvi-2) and in references (1), LINE 3;  (3), LINE I1;  and (21), LINE RC12.
b  TACTCA or any of its 18 single base substituted variants.
c  $P_n$ is the chance probability of finding 2 occurrences of TACTCA, or any of its variants, in the 76 bp of target site that flanks LINE Mlvi-2 (40 bp on the left, 36 bp on the right), or the 3 occurrences of TACTCA, or variants, in the 78 bp of target site that flanks LINE 3 (40 bp on the left, 38 bp on the right).
d  This value is $P_2$.

A-rich region.  The remainder of the sequenced region of non-LINE DNA beyond LINE I1 (164 bp) did not contain PQ stretches.  However, the corresponding region beyond LINE RC12 contained several, two of which are shown.  In contrast to the sequences in Fig. 3A, these stretches were not simple multiples of a single PQ pair.

For 4 of the 5 sequences shown in Fig. 3, a version (0 or 1 mismatch) of the LINE hexamer, TACTCA, is found in the target site DNA within 41 bp of the right end of the LINE.  In the case of RC12, this distance is 73 bp.  Each of these righthand flanking sequences is considerably more (A+T)-rich than total rat DNA [0.58 (A+T) (5)];  see the right side of Table 1.  Furthermore, the probability that a TACTCA or any of its variants occurred in all 5 of the flanking sequences by chance is about $1.7 \times 10^{-5}$.  This value is the product of the chance probabilities for 1 occurrence (or 2 in the case of LINE 3) of TACTCA or its variants in each righthand flanking sequence (see the right side of Table 1 and "Materials and Methods").  The consensus hexamer sequence found is NPCTNA, where N is any nucleotide and P is purine (see Fig. 3).

The left part of Table 1 shows the (A+T) content and the chance probability of finding a single occurrence of any of the hexamers in the 40 bp of

target site that flanks the left side of the various LINE members. The (A+T) content of each left flanking sequence was less than that of the corresponding right flanking sequence. However, the left flanking sequences of LINE Mlvi-2 and LINE 3 were each more (A+T)-rich than total rat DNA, and each contained one occurrence of a hexamer variant. The chance probabilities for the 2 occurrences of TACTCA or its variants in the ≅ 80 bp that flank LINE Mlvi-2, and for the 3 occurrences of the hexamers in the ≅ 80 bp that flank LINE 3 are 0.09 and 0.03, respectively (last column of Table 1). Taken together, these and the above results indicate that the LINE members shown in Fig. 3 inserted just to the left of, or into, an (A+T)-rich sequence and that the occurrence of at least one copy of TACTCA or its variants in these sequences is most unlikely to be due to chance.

Fig. 3 also reveals several noteworthy features of the A-rich right ends of the rat LINE members: First, they vary in length and, except for LINE RC12, each contains direct repeats. Second, the A-rich end of LINE 3 is not only considerably longer than the others but also has a more complex base composition and contains the hexanucleotide, TACTCA. Comparison of this portion of the A-rich sequence with the relevant portion of the LINE sequence that contains the LINE TACTCA [at position 1 in Fig. 3, at position 6629 in LINE 3 (1)] suggests that the TACTCA-containing A-rich right end of LINE 3 is not the result of a simple duplication of the LINE sequence. An alternative explanation is that the A-rich end of LINE 3 includes the target site from a previous transposition, which in turn would mean that LINE 3 has transposed at least twice.


## DISCUSSION

The empty Mlvi-2 target site is (A+T)-rich, especially that region which, in the LINE-containing allele, flanks the right end of the LINE (see Table 1). This region also contains a TACTCA, and just to the right of it is a stretch of simple PQ. The only major change in target site DNA that is ascribable to the LINE insertion is the duplication of a 14-bp sequence which flanks the inserted LINE.

As Table 1 shows, an (A+T)-rich stretch that contains at least one occurrence of a singly-substituted variant of TACTCA flanks the right end of 4 other LINE members. In all cases the (A+T) content is significantly higher than that of total rat DNA, and in each case where we could determine it the duplicated target site sequence is at least 40% A and contains one or more stretches of A. The occurrence of TACTCA or its variants in all of these sequences is highly statistically significant (P ≅ 1.7 x $10^{-5}$). It is

important to stress that only 2 of these 5 LINE-containing sites (LINEs 3 and 4) were selected directly for LINE sequences (1). The other 3 were originally isolated because of interest in their single copy DNA. Only after additional studies were LINE members detected (2, 3, 21).

Three of the 5 LINE target sites contain a stretch of simple PQ at a variable distance to the right of LINE insertion (Fig. 3A). Therefore, an (A+T)-rich stretch that contains a TACTCA or singly-substituted variant thereof, followed by a simple PQ stretch, typifies at least one class of LINE target sites. We roughly estimated the number of such sites in the rat genome as follows: Mammalian genomes contain from about $0.1 \times 10^6$ (22) to 1 $\times 10^6$ (20) copies of $(GT)_n$, where n is 25 (22) or 8 (20). GC stretches are far fewer (22), and assuming that AT stretches, which were not determined, are as abundant as GT stretches, then the rat genome could contain from $0.2 \times 10^6$ to $2 \times 10^6$ copies of simple PQ stretches. Only some of these would be in a region of DNA that is both (A+T)-rich and contains a TACTCA or version thereof. Since (A+T)-rich regions are likely to have stretches of A, we estimated the number of such regions by calculating the probability of finding an $(A)_5$ sequence and a TACTCA or version thereof by chance in 300 bp of DNA having the base composition of total rat DNA. This value is 0.135 [i.e., the probability of an $(A)_5$, 0.3 x the probability of the hexamer, 0.45]. Therefore, if simple (PQ) stretches are dispersed randomly in the genome, then there is the chance probability of 0.135 that some would be within 300 bp of such an (A+T)-rich region. Therefore, the rat genome could contain as many as $0.135 \times 2 \times 10^6$ or about 270,000 such sites. If LINE transposition were completely random, then we calculate that given the size of the rat genome ($\cong 3 \times 10^9$ bp) and the number of LINE members ($\cong 40,000$), the chance of finding the right end of a LINE member at a target site to be in the order of $10^{-9}$ [i.e., $(2.7 \times 10^5 / 3 \times 10^9) \times (4 \times 10^4 / 3 \times 10^9)$]. Even if this estimate is off by several orders of magnitude, the insertion of 3 of 5 LINE members at such sites could not have been random.

The fact that rat LINE members do not insert randomly in the genome suggests that target site sequences play a role in LINE transposition. Although our current results do not warrant further speculations, we note the following: First, both the target site and the right end of LINE members contain TACTCA, or a variant hexamer in the case of some target sites, and stretches of A's (Fig. 3). Second, stretches of PQ sequence, even short imperfect ones such as those in the LINE RC12 target site (Fig. 3B), have the potential to form Z-DNA (23, 24) as well as other non-B DNA structures, such as a cruciform, perhaps via a Z-DNA intermediate (25). Such non-B DNA

conformations, or the transitions to and from them, may be involved in various regulatory or recombinational events (23, 25). In particular, the rec-1 protein of Ustilago binds tightly to Z-DNA, which may be an intermediate in the recombination promoted by this protein (26). Furthermore, PQ stretches have been involved in or are near two SV40 integration sites in rat DNA (27, 28).

The presence of direct repeats in the A-rich end of rat LINE members suggests that this end might have originated as a "true" LINE sequence [e.g., AAT(A)n] that was subsequently modified by the processes known to affect simple sequences such as this [see Results and (20)]. Furthermore, as might be the case for LINE 3, an A-rich end could also expand by acquiring the A-rich target site from a previous transposition. These possibilities contrast with the idea that the A-rich right end of LINE members is the relic of a polyadenylated tail of a reverse transcribed (retrotranscribed) LINE transcript (9-11, 29). Although retroposition (9) of LINE members could account for some instances of LINE transposition, other mechanisms are also possible (1, 10). Furthermore, the idea that a transposed LINE member is analogous to a retroposed pseudogene and is therefore a passive, inert "pseudomember" of the LINE family (9) is at variance with the possibility that LINE 3 transposed more than once.

The above possibility, as well as the identification of certain sequence elements in rat LINE target sites, suggests certain experimental approaches to the problem of LINE amplification and transposition. For example, proteins involved in these processes may be among those that bind to Z-DNA (30, 31). Furthermore, the positioning of empty or LINE-containing target sites in the appropriate regions of cloned genes that permit selection either for the loss or recovery of activity could provide an assay for LINE insertion or excision.

*Present address: Department of Microbiology, University of Maryland, College Park, MD 20742, USA

REFERENCES
1.  D'Ambrosio, E., Waitzkin, S. D., Witney, F. R., Salemme, A. and Furano, A. V. (1986) Mol. Cell. Biol. 6, 411-424.
2.  Economou-Pachnis, A., Lohse, M. A., Furano, A. V. and Tsichlis, P. N. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 2857-2861.
3.  Lakshmikumaran, M. S., D'Ambrosio, E., Laimins, L. A., Lin, D. T. and A. V. Furano (1985) Mol. Cell. Biol. 5, 2197-2203.
4.  Tsichlis, P. N., Strauss, P. G. and Loshe, M. A. (1985) J. Virol. 56, 258-267.
5.  Shapiro, H. S. (1976) In Fasman, G. D. (ed), Handbook of Biochemistry and Molecular Biology:  Nucleic Acids, CRC Press, Cleveland, Ohio, Vol. 2, pp. 241-311.
6.  Singer, M. F. (1982) Int. Rev. Cytol. 76, 67-112.
7.  Voliva, C. F., Jahn, C. L., Comer, M. B., Hutchison, C. A. III and Edgell, M. H. (1983) Nucleic Acids Res. 11, 8847-8859.
8.  Britten, R. J. and Kohne, D. E. (1968) Science 161, 529-540.
9.  Rogers, J. H. (1985) Int. Rev. Cytol. 93, 187-279.
10. Singer, M. F. and Skowronski, (1985) J. Trends Biochem. Sci. 10, 119-122.
11. Fanning, T. G. (1983) Nucleic Acids Res. 11, 5073-5091.
12. Lehrman, M. A., Schneider, W. J., Südhof, T. C., Brown, M. S., Goldstein, J. L. and Russell, D. W. (1985) Science 227, 140-146.
13. Vieira, J. and Messing, J. (1982) Gene 19, 259-268.
14. Bencini, D. A., O'Donavan, G. A. and Wild, J. R. (1984) Biotechniques 2, 4-5.
15. Maxam, A. M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
16. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467.
17. Norrander, J., Kempe, T. and Messing, J. (1983) Gene 26, 101-106.
18. Henikoff, S. (1984) Gene 28, 351-359.
19. Handbook of Chemistry and Physics (1964) Weast, R. C., Selby, S. M. and Hodgman, C. D. (eds), CRC Press, Cleveland, Ohio, p. A157.
20. Tautz, D. and Renz, M. (1984) Nucleic Acids Res. 12, 4127-4138.
21. Scarpulla, R. C. (1985) Nucleic Acids Res. 13, 763-775.
22. Hamada, H., Petrino, M. G. and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 6465-6469.
23. Rich, A., Nordheim, A. and Wang, A. H.-J. (1984) Annu. Rev. Biochem. 53, 791-846.
24. Feigon, J., Wang, A. H.-J., Van Der Marel, G. A., Van Boom, J. H. and Rich, A. (1985) Science 230, 82-84.
25. Haniford, D. B. and Pulleyblank, D. E. (1985) Nucleic Acids Res. 13, 4343-4363.
26. Kmiec, E. B., Angelides, K. J. and Holloman, W. K. (1985) Cell 40, 139-145.
27. Bullock, P., Forrester, W. and Botchan, M. (1984) J. Mol. Biol. 174, 55-84.
28. Stringer, J. R. (1982) Nature 296, 363-366.
29. Voliva, C. F., Martin, S. L., Hutchison III, C. A. and Edgell, M. H. (1984) J. Mol. Biol. 178, 795-813.
30. Nordheim, A., Tesser, P., Azorin, F., Kwon, Y. H., Müller, A. and Rich, A. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 7729-7733.
31. Azorin, F. and Rich, A. (1985) Cell 41, 365-374.
32. Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H. and Hutchison, C. A. III (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 2308-2312.