Volume 16 Number 3 1988 Nucleic Acids Research

Human transposon-like elements insert at a preferred target site: evidence for a retrovirally mediated process

Niren Deka, Cary R.Willard, Elsie Wong and Carl W.Schmid

Department of Chemistry, University of California, Davis, CA 95616, USA

Received October 13, 1987; Revised and Accepted December 29, 1987

ABSTRACT

Members of the human transposon-like family of repetitive sequences (called THE 1 repeats) like many other repetitive DNA sequences are flanked by short direct repeats. Comparison of the base sequences of twelve examples of these flanking direct repeats indicates that THE 1 repeats insert into a preferred genomic target site. In one case, we have identified the sequence of an empty site into which a THE 1 element inserted. The sequence of this empty site and sequences of truncated THE 1 LTRs are consistent with a retroviral mechanism for the insertion of THE 1 elements. Truncated transposon structures illustrate for the first time that intermediate structures of retrotransposition may also be integrated into the genome.

INTRODUCTION

Many eukaryotic DNAs contain proretroviral-like repetitive elements such as Ty elements in yeast, copia-like elements in Drosophila, IAP repeats in mouse and THE 1 repeats in human (1,2,3,4). The biochemical steps involved in the transposition of retroviral RNA sequences give rise to a complementary DNA, which is subsequently ligated into a circular DNA intermediate (5,6,7). The ligated junction of the circular DNA intermediate is sealed by formation of a novel tetranucleotide sequence, TTAA, which is removed upon insertion of the proretroviral DNA sequence (6,7). The circular intermediate is inserted into the genome by a virally encoded integrase activity which evidently recognizes the dyad symmetry ...CATTAATG... present at the ligated junction sequence (7,8). Retroviral gene products specifically cleave this sequence in vitro (6,7). Thus, the integration system giving rise to newly inserted proretroviral sequences is potentially sequence specific.

In general, proretroviral sequences insert in a sequence independent manner (8,9). However, yeast Ty elements are found to insert into A rich regions (10) and 17.6 elements in Drosophila are found to insert into a preferred target site resembling the promoter sequence TATAA (11). The target sites for the transposition of rat long interspersed repeats (LINES) are also found to be preferentially A rich (12). Preliminary sequence results suggested that THE 1 repetitive elements in human DNA might also insert in a sequence specific manner (4).

The 2.3 kb consensus sequence of THE 1 repeats is similar to that of proretroviral structures in that most members are flanked by two 350 bp LTRs (4). The LTR sequences begin with 5' TG... and end with ...CA 3' as is typical of most proretroviral sequences (4). In addition to the approximately 10,000 full length copies of THE 1, human DNA also contains about 10,000 solitary THE 1 LTRs. Except for this overall structural similarity THE 1 repeats are unrelated to any known retroviral sequence. Here we examine the direct repeats flanking several full length THE 1 sequences and several solitary LTRs.

MATERIALS AND METHODS

Several lambda phage genomic clones from the Maniatis human library were isolated using THE 1 specific probes (4). Genomic clones T^+ and T^- were isolated using unique sequence flanking a THE 1 containing cDNA α (15). Subclones in PUC were constructed and grown by standard methods (13). M13 subclones for sequencing were constructed by insertion into the universal cloning sites of M13 (14). Four of the eleven cloned sequences analyzed here have been previously published (4,20,23). As we are interested in only the direct repeat in the present project M13 sequencing clones were constructed very near the ends of the LTRs to scrutinize the direct repeats. Three of the remaining clones were sequenced in both strands in the vicinity of the direct repeats; the remaining examples were sequenced in only one strand with multiple readings.

RESULTS

Many THE 1 short direct repeats share a consensus sequence

λ genomic clones containing THE 1 repeats or solitary LTRs were selected from the Maniatis human library. The THE 1 sequences in several of these clones were entirely determined but exhibited no noteworthy differences compared to the previously published sequence of THE 1-A (4). For this reason, we shall present here only the sequences flanking THE 1 elements. These data are compiled in Figure 1A.

With three exceptions, THE 1-B, C, and D, the sequences compiled in Figure 1 have complete THE 1 LTRs. THE 1-B, C and D all have truncated 3' LTRs (Figure 2). THE 1-C is exceptional in that its 3' LTR is truncated by a poly A stretch corresponding to the 3' end of a THE 1 mRNA-like sequence (15). Furthermore, the 5' LTR of THE 1-C is also truncated. When the 5' and 3' THE 1-C LTR sequences are overlapped the following three sequence features can be discerned; a) a 36 nt long sequence which is unique to the left LTR "U5"; b) a 98 nt long sequence repeated in both the LTRs "R" and c) a 260 nt long sequence which is unique to the right LTR "U3" (Figure 2). These U3, R and U5 sequence features are well defined in the retroviral long terminal repeats as well as the copia and Ty LTRs which share a common structure of 5' U3 RU5 3' (5, 16, 19). During the transcription of retroviral-like elements, the boundary between the U3 and R region in the left LTR defines the transcription start site and the R U5 adjacent region provides the polyadenylation signal (5). Consequently, these retroviral mRNAs share a common structure of 5' **RU5 coding region U3R poly A** 3' (5,16,17). THE 1-C evidently corresponds to a THE 1 RNA structure and resembles retroviral-like copia and Ty RNAs. In agreement with this interpretation, the 3' LTRs of the THE 1-B and THE 1-D are both truncated almost exactly at the proposed boundary of the U3 and R regions of the LTR (Figure 2). Presumably, these incomplete THE 1 members arose by a defective reverse transcription of their RNA intermediates.

Several of the direct repeats flanking THE 1 elements resemble the THE 1-A direct repeat 5'CAGATAC 3' (Figure 1A).

A dr		THE	dr
CTCAGATAC	TG	THE 1 A	CA GATACCGAG
12345			
GTATCATAC	TG	THE 1 B truncated	ATATGAGG
GTAAGATAT	TG	THE 1 D truncated	AGCTGTGC
CAAAGGCCC	TG	THE 1 F	CA GCCATGCCAT
ATTTAGGC	TG	THE 1 J	CA CAGGCTTT
GAAAGATGC	TG	THE 1 T^+	CA GATGCTGAAGAA
TCTACCTCAC	truncated	THE 1 C Poly A truncated	GTGTTCAAGT
		LTR	
TAATGACCAC	TG	Sol A	CA CCAGCT
GGAACTGAA	GG	Sol B	CA CTCT
CATATCATT	TG	Sol O_4	CA TCATTCGT
CCGTGATTAC	TG	Sol O_5	CA G-TTACTATGA
TGAGACCCAC	TG	Sol O(ADA)	CA G-CC-ACCT

B

Position	Frequency				
	1	2	3	4	5
Base					
A	2	5	1	8	1
G	4	1	2	2	-
C	4	3	4	1	9
T	2	3	5	1	2
Consensus	**g** **c**	**a**	**y**	**A**	**C**

Fig. 1 A. **Short direct repeat sequences flanking THE 1 and solitary LTR members.** The underlined sequences denote the short direct repeats at the site of insertion of individual THE 1 and the solitary LTR members. TG and CA demarcate the boundaries of THE 1 and the solitary LTR members. The numbers 1 through 5 underneath the short direct repeat of THE 1-A indicates the preferred target site of THE 1 as shown in Fig. 1B. Asterisks indicate the imperfect base in the short direct repeats. THE 1-B, 1-D and 1-C are truncated in their 5' or 3' LTRs.

Fig. 1 B. **Preferred insertion site of THE 1 repeat.** For explicitness, we show only the left direct repeat which is well defined as immediately flanking the TG which initiates the THE 1 element. The frequency of occurrence of individual bases A, G, C, and T are compiled under the number headings 1 to 5. These numbers 1 through 5 refer to the positions of bases in the short direct repeat sequence as indicated underneath the short direct repeat flanking THE 1-A in Fig. 1 A. The consensus insertion sequence of AC is derived on the basis of highest frequency of occurence of individual bases without including possible single nucleotide deletions or substitutions. More extended preference could include a pyrimidine at position 3 and an adenine at position 2 such that the preferred target might be represented as ayAC.

However, it is ambiguous whether the dinucleotide CA should be considered as part of THE 1-A's direct repeat as it also forms the 3' terminus of the right LTR (Figure 1A). For this reason,

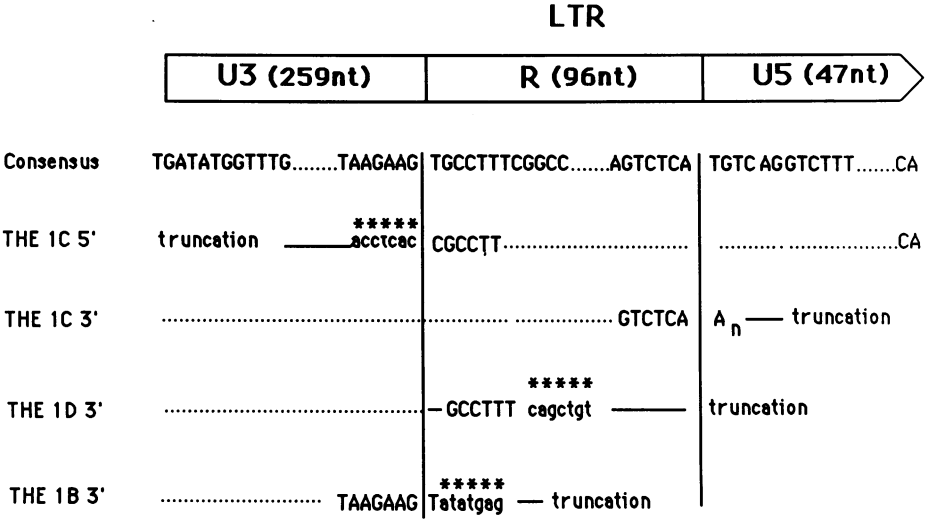

Fig. 2. **Sequence alignments at the boundaries of U3, R and U5 regions for THE 1 LTRs.** The consensus sequence of a THE 1 LTR at the boundaries of U3, R and U5 regions is shown. Sites of truncation are indicated for the 5' and the 3' LTRs of THE 1-C and for 3' LTRs of THE 1-D and 1-B. U3, R and U5 represent the unique 3' region, the repeat region and a unique 5' region respectively of a THE 1 LTR. The dotted lines indicate the conserved sequence and the asterisk (*) indicates non-homologous sequences compared to the consensus LTR sequence.

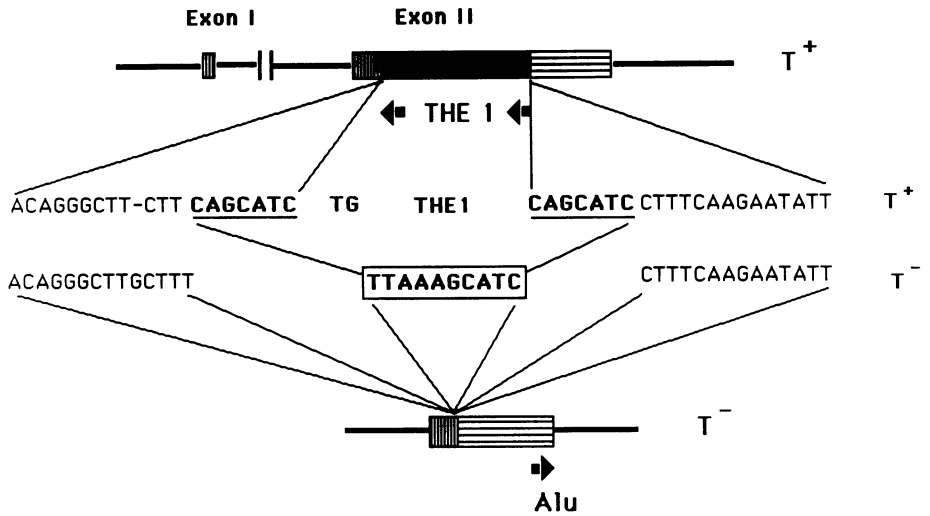

Fig. 3. **Sequence of an empty site and resulting direct repeat upon insertion of a THE 1 repeat.** T^+ and T^- are two genomic loci related to each other through their homology in the 5' and 3' unique regions (see text). The empty site is indicated within the boxed area. The 7 nt long short direct repeat of 5' CAGCATC 3' flanking the THE 1 is underlined. The complement of this short direct repeat e.g. 5' GATGCTG 3' is shown as the short direct repeat of THE 1 T^+ in Fig. 1 A. The dash indicates an insertion/deletion of a single nucleotide.

we shall consider only 5' GATAC 3' as being the direct repeat of THE 1-A. Allowing for transition mutations and deletion/ insertion variants of the THE 1-A direct repeats can be recognized as the direct repeats flanking THE 1-B, 1-C, 1-D, T^+, Sol A, O_5 and O(ADA). The short direct repeats flanking solitary O(ADA) has been reported elsewhere (20). The most prevalent nucleotides at positions 4 and 5 are AC (Fig. 1B). Additional preference for an A at position 2 and a pyrimidine at position 3 is less convincing, thus we designate the preferred target as resembling ayAC where small letters suggest an extended preference (Figure 1B). This sequence preference is similar to the sequence specific insertion of Drosophila 17.6 into the tetranucleotide 5'...ATAT... 3' (11).

Comparison of an empty target site to the resulting direct repeats

We wish to examine an empty target site to understand better the sequence preference of THE 1 insertion. Previously, we

identified a cDNA clone to a transcript containing a full length THE 1 member, cDNA α (15). Unique sequences within this cDNA clone hybridize to two distinct loci in the human genome in all individuals tested. These loci have been isolated as two λ genomic clones called T^+ and T^-. One genomic clone, T^+, contains an inserted THE 1 repeat whereas the THE 1 repeat is absent in T^- (Figure 3). Base sequence data not shown here reveal that T^+ is the gene encoding cDNA α whereas T^- is a processed RNA pseudogene which is ancestrally related to cDNAα. [As evidence for these interpretations, T^+ has recognizable promoter sequences and an intron with consensus splice sites relative to the cDNAα sequence. In contrast, the T^- sequence has a number of point mutations compared to cDNAα. Unlike cDNAα, THE 1 is absent from T^- but an Alu repeat has been inserted into T^-. Further the T^- region of homology to cDNAα is polyadenylated and is flanked by short direct repeats as is typical of processed RNA pseudogenes].

The direct repeats ...AGCATC... flanking THE 1 in T^+ identify the empty target site in T^- which is duplicated upon THE 1 insertion (Figure 3). It should be noted that the THE 1 element in T^+ is oriented in opposite direction compared to other members such as THE 1A, 1-B etc., hence the direct repeat is properly read as ...GATGCT (see in Figure 1A). Immediately adjacent to the empty target site is the tetranucleotide ...TTAA... which was apparently removed upon insertion of THE 1. Because the T^+ and T^- loci are paralogously diverging sequences, it is also formally possible that the tetranucleotide TTAA was independently inserted at this site in the T^- locus. We discount this possibility as being improbable because of the otherwise excellent sequence similarity of the T^+ and T^- loci (Figure 3).

DISCUSSION

The insertion of THE 1 members into a preferred target sequence is similar to the preferential insertion of yeast Ty and Drosophila 17.6 transposons (10,11). Since most proretroviral/ transposon sequences insert without a recognizable sequence preference, these three elements are exceptional. The biochemical basis of this preference is not known. However, sequence specificity does occur in at least one intermediate step

in transposition; cleavage of the circular DNA intermediate occurs precisely at the sequence joining the two LTRs 5'...CATTAATG..3' (6,7). THE 1 elements are present as extrachromosomal circular DNAs but the known structures of these circular DNAs do not correspond to that expected for transposition intermediates (4,18). The genomic target site, like the junction of the circular DNA intermediate, must also be susceptible to nuclease cleavage and like the circular junction might be preferentially cleaved at particular sequences. The presence of the tetranucleotide TTAA at the empty target site and its removal upon THE 1 insertion evokes the properties of a retrovirally encoded insertion factor.

Truncated THE 1 elements are also consistent with a retrovirally directed dispersion of THE 1 members. The ends of the LTR sequences are known to be important for integration (6). Thus, it is surprising that truncated THE 1 members insert with the same target preference proposed for complete THE 1 elements. Presumably, the same trans-acting factors recognize both full length and truncated THE 1 members. The retroposition of mRNA can be directed by retrovirus (21). It is noteworthy that THE 1C resembles a classic retroposon (i.e. processed RNA pseudogene) and has the structure predicted for the intermediate retroviral mRNA [5' **RU5 coding region - U3R poly A** 3'] (5,22, Figure 2). Similarly, the truncated members THE 1B and THE 1D also appear to be aberrant proretroviral intermediates (Figure 2). Processed mRNA pseudogenes, retroposons, and proretroviral transposons are end products of the reverse flow of genetic information and thus require many of the same catalytic pathways. The THE 1 family is unusual in that its membership includes both retroposons and proretroviral transposons.

Acknowledgement

This research was supported in part by USPHS Grant GM 21346.

References

1. Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979) Cell **16**, 739-751.
2. Rubin, G.M. (1983) Chapter 8 "Dispersed repetitive DNAs in Drosophila". In Shapiro J.A. (ed), Mobile Genetic Elements, pp. 329-361. Academic Press, Orlando, Florida.

3. Kuff, E.L., Feenstra, A., Lueders, K. Smith, L., Hawley, R., Hozumi, N. and Schulman, M. (1983) Proc.Natl.Acad.Sci.USA **80**, 1992-1996.
4. Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G. Kadyk, L. and Leinwand, L. (1985) Nature **316**, 359-361.
5. Varmus, H.E. (1982) Science **216**, 812-820.
6. Colicelli, J. and Goff, S.P. (1985) Cell **42**, 573-580.
7. Panganiban, A.T. and Temin, H.M. (1984) Cell **36**, 673-679.
8. Grandgenett, D.P. and Vora, A.C. (1985) Nucl.Acids Res. **13**, 6205-6221.
9. Brown, P.O., Bowerman, B., Varmus, H.E. and Bishop, J.M. (1987) Cell **49**, 347-356.
10. Eibel, H. and Philippsen, P. (1984) Nature **307**, 386-388.
11. Inouye, S., Yuki, S. and Saigo, K. (1984) Nature **310**, 332-333.
12. Furano, A.V., Somerville, C.C., Tsichlis, P.N. and D'Ambrosio, E. (1986) Nucl.Acids Res. **14**, 3717-3727.
13. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1983) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Lab., NY
14. Messing, J. Crea, R. and Beeburg, P.H. (1981) Nucl.Acids Res. **9**, 309-323.
15. Paulson, K.E., Matera, A.G., Deka, N. and Schmid, C.W. (1987) Nucl.Acids Res. **15**, 5199-5215.
16. Elder, R.T., Loh, E.Y. and Davis, R.W. (1983) Proc.Natl. Acad.Sci. USA **80**, 2432-2436.
17. Flavell, A.J., Levis, R., Simon, M. and Rubin G.M. (1981) Nucl.Acids Res. **9**, 6279-6291.
18. Misra, R., Shih, A., Rush, M., Wong, E. and Schmid, C.W. (1987) J.Mol.Biol. **196**, 233-243.
19. Arkhipova, I.R., Mazo, A.M., Cherkasova, V.A., Gorelova, T.V., Schuppe, N.G. and Ilyin, Y.V. (1986) Cell **44**, 555-563.
20. Wiginton, D.A., Kaplan, D.J., States, J.C., Akeson, A.J., Perme, C.M., Bilyk, I.J., Vaughn, A.J., Lattier, D.J. and Hutton, J.J. (1986) Biochemistry **25**, 8234-8244.
21. Linial, M. (1987) Cell **49**, 93-102.
22. Temin, H.M. (1981) Cell **27**, 1-3.
23. Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984) Nucleic Acids Res. **12**, 2669-2690.