J. Mol. Biol. (1995) **247**, 49–59

JMB

Characterization of a Non-long Terminal Repeat Retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: Homology of the First ORF with the Ape Family of DNA Repair Enzymes

Francisco Martín[1], Concepción Marañón[1], Mónica Olivares[1] Carlos Alonso[2] and Manuel C. López[1]*

[1]*Departamento de Biología Molecular, Instituto de Parasitología y Biomedicina ''López Neyra'', C.S.I.C. Calle Ventanilla no. 11, 18001 Granada, Spain*

[2]*Centro de Biología Molecular ''Severo Ochoa'' C.S.I.C.-U.A.M. Cantoblanco 28049 Madrid, Spain*

**Corresponding author*

In the present paper we describe the characterization of a *Trypanosoma cruzi* cDNA (L1Tc) corresponding to a transcript from a new long terminal repeat (LTR) retrotransposon. This element is present in a high-copy number, and is found dispersed throughout the *T. cruzi* genome. Northern analysis shows an abundant expression of L1Tc-related sequences with a major band of about 5 kb. The transcript has at its 3' end a fragment of a highly repetitive DNA sequence (E12A), at its 5' end a ribosomal mobile element-like sequence and three putative open reading frames (ORF) in different frames. The ORF2 codes for a protein which has significant homology with the retrotranscriptase-related sequences from non-LTR retrotransposons containing the seven domains present in all the retrotranscriptase and retrotranscriptase-related proteins. The ORF3 codes for a gag-like protein showing unusual cysteine motifs present in all non-LTR trypanosomatid elements, similar to the C_2H_2 zinc finger family of transcription factors. Interestingly, ORF1 codes for a protein with significant homology to the major human AP endonuclease protein, and maintains in similar positions most of the amino acid domains described for all the Ape family of proteins. The presence of Ape-related sequences, described for the first time in a non-LTR retrotransposon (L1Tc), may have functional relevance for these types of elements.

Keywords: transposable element; reverse transcription; repair enzymes; ribosomal mobile elements like; *Trypanosoma cruzi*

Introduction

The highly repetitive DNA sequences of most eukaryotic cells represent a large fraction of their nuclear DNA. These highly repetitive DNA sequences can be classified into the satellite (single DNA) class, formed by tandemly arranged fragments of short oligonucleotides, and into the interspersed class represented by the LINE and SINE families (Singer, 1982). The evolutionary origin of these sequences is uncertain, but it has been suggested that the LINEs and SINEs may correspond to partial or complete DNA copies of cellular RNA transcripts (Weiner *et al*., 1986), and that they may represent a class of transposable elements, recently named as the non-LTR retrotransposons (Xiong & Eickbush, 1988). These non-LTR retrotransposons have been identified in a wide variety of eukaryotic organisms and may constitute as much as 5% of the genome (Singer & Skowronski, 1985; Fawcett *et al*., 1986; Hutchinson *et al*., 1989; Leeton & Smyth, 1993). In certain organisms, however, some non-LTR retrotransposons are present in a low copy number (Xiong & Eickbush, 1988; Morse *et al*., 1988; Aksoy *et al*., 1990; Gabriel *et al*., 1990; Villanueva *et al*., 1991). Only recently, it has been demonstrated that some non-LTR retrotransposons are capable of transposition, and that this transposition is mediated *via* an RNA intermediate (Eickbush, 1992). Most non-LTR retrotransposons display two ORFs in different reading frames which overlap for a short distance. These ORFs encode enzymes which could be involved in their own transposition (Eickbush, 1992). ORF1 contains cysteine motifs similar to those of the

Abbreviations used: LTR, long-terminal repeat; ORF, open reading frame; RT, reverse transcriptase; RIME, ribosome mobile element.

0022–2836/95/110049–11 $08.00/0

retroviral *gag* genes. ORF2 has sequence similarity to the retroviral *pol* genes, particularly in the 300 amino acid domain of the reverse transcriptase (RT).

In the Trypanosomatidae, the search for highly repetitive DNA sequences has been a major research goal for many laboratories, because it has been postulated that these sequences may play an important role in genome structure and expression, and because they may also be used for highly sensitive parasite detection (González *et al*., 1984) and strain classification (Requena *et al*., 1992). It is interesting to note that two of the trypanosomatid repetitive DNA sequences have been described as non-LTR retrotransposons (Murphy *et al*., 1987; Kimmel *et al*., 1987). Other trypanosomatid non-LTR retrotransposons, described as site specific, are present in a low copy number (Aksoy *et al*., 1990; Gabriel *et al.*, 1990; Villanueva *et al*., 1991).

In the course of the analysis of highly repetitive sequences from *Trypanosoma cruzi* nuclear DNA it has been found that a fragment of a repetitive element, named E12A, is present in a 5 kb long cDNA highly represented in poly $(A)^+$ RNA (Requena *et al*., 1994). In the present paper we describe the characterization of the E12A containing a 5 kb long transcript that includes a ribosomal mobile element (RIME)-like sequence at its 5' end. This transcript is a high copy number non-LTR retrotransposon which contains three non-overlapping ORFs in different frames. ORF2 and 3 show homology with the *pol*- and *gag*-encoded proteins of non-LTR retrotransposons. Interestingly, ORF1 showed significant homology with the Ape family proteins (Demple *et al*., 1991).

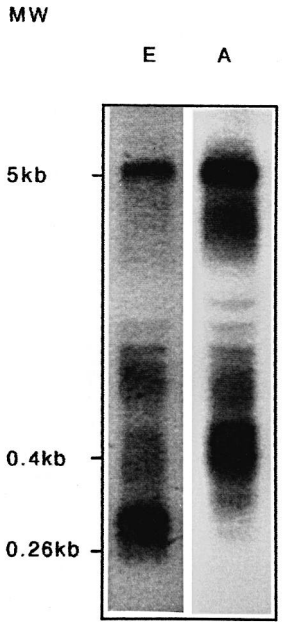

Figure 1. Northern blot analysis of *T. cruzi* RNA. A 2 μg sample of poly$(A)^+$ RNA fractionated on a 1% agarose–formaldehyde gel and blotted into a nylon membrane was probed with the *Eco*RI–*Eco*RI 5'-end (E) and the *Aat*II-*Aat*II 3'-end (A) fragments of L1Tc. The numbers at the left-hand side indicate the size of the most intensively labeled mRNA bands.

Results

L1Tc cDNA shows homology at its 5' end with the RIME sequence, and contains three non-overlapping ORFs

It has been previously reported that a highly repetitive element (E12) of *T. cruzi* chromosomal DNA is present in several RNA transcripts of different lengths (Requena *et al*., 1994). The most intensively labeled RNA band corresponds to a transcript of approximately 5 kb. In order to isolate transcripts containing E12, a cDNA expression library of *T. cruzi* epimastigotes was probed with the E12 repetitive element. Several positive hybridization clones of different were isolated. One of the clones (pSPFM55) containing a 5.0 kb long insert was chosen for analysis. The cDNA insert of the clone was called L1Tc. The hybridization to *T. cruzi* poly $(A)^+$ RNA of the 5' end (*Eco*RI-*Eco*RI) and the 3' end (*Aat*II-*Aat*II) fragments of L1Tc indicated that the L1Tc was present in the 5.0 kb long RNA band (Figure 1). The complete nucleotide sequence and the deduced amino acid sequence from L1Tc are shown in Figure 2. The analysis of the nucleotide sequence revealed that a fragment of E12, called E12A, was present in the 3' end of the transcript, and that the 5' end of L1Tc also showed significative homologies with RIME and RIME-like also found in retrotransposons from *Trypanosoma brucei*. In fact, the nucleotide sequence from the RIME of the VSG gene expression site (Tbesag) (Pays *et al*., 1989) shows a 69% identity with 127 nucleotides of L1Tc. This homology extends from nt 5264 to nt 5391 in Tbesag, and from nt 65 to nt 192 in L1Tc. There is, moreover, significant homology of the 5' end of L1Tc with the RIME DNA sequence from Tbbs12 (Hobbs & Boothroyd, 1990), Tbtrs16 (Murphy *et al*., 1987), Tbingi (Kimmel *et al*., 1987), Tbtubb3 (Affolter *et al*., 1989), Tbvsga (Hasan *et al*., 1984) and Trrgrime (Hasan *et al*., 1982) of *T. brucei*. Also, the *T. cruzi* Tcaad (Baschiazo *et al*., 1992) and Tcanta (Bontempi *et al*., 1993) sequences show high homology (72.8% and 74.2% of identity, respectively) with the 5' end of L1Tc. The homology is revealed with the reverse and complementary strands of 213 nt from the untranslated 5' region of Tcaad and 66 nt from the 3' region of Tcanta. The DNA sequence between nt 119 and nt 148 of L1Tc is 90 to 95% conserved in all RIME and RIME-like sequences.

The analysis of the nucleotide sequence of L1Tc cDNA (Figure 2) showed three ORFs in different reading frames. The first ORF (frame 1) began at nt 102 and ended at the TAA termination codon in nt 1228. The second ORF (frame 2) extended from nt 1799 to the TAA termination codon at nt 3623. The third ORF (frame 3) began at nt 3993 and continued to the TAG termination codon at nt 4965. The predicted amino acid sequence of each ORF was named L1Tca, L1Tcb and L1Tcc, respectively.

L1Tca (ORF1) showed homology with the Ape family of DNA repair enzymes and with the *pol* gene N terminus of several non-LTR retrotransposons

The comparison of the predicted amino acid sequence of L1Tca with proteins available in the sequence database using the BLASTP program revealed homologies with the Ap1–human protein, which belongs to a family of repair enzymes denominated as the Ape family (Demple *et al*., 1991). The analysis of the homology between the sequence of L1Tca and the proteins of the Ape family Ap1-human (Demple *et al*., 1991), Rrp1-drome (Sanders *et al*., 1991), Ex3-ecoli (Saporito *et al*., 1988) and ExoA-strpn (Puyet *et al*., 1989), using the BESTFIT program, showed the presence of various levels of homology between these proteins. L1Tca showed the greatest homology with the Ap1-human protein with a 20.3% identity in a 212 amino acid fragment and a *Z* score of 10.1. This *Z* value revealed that the similarity between the Ap1-human protein and L1Tca is biologically significant (Doolittle, 1981). The PILEUP program was used to align some of the proteins of the Ape family with L1Tca (Figure 3). Several highly conserved regions were found to have, in similar positions, most of the amino acid domains described for all the Ape family proteins. The most highly conserved domains were found in the amino acid positions 124 to 133, 154 to 172 and 228 to 242.

The TFASTA analysis showed similarities between L1Tca, the protein coded in ORF1 of the *T. brucei* ingi-3 element and the *pol* protein N terminus from some non-LTR retrotransposons. The BESTFIT program showed a 30.8% of identity of L1Tca with the polypeptide coded by the ORF1 of the ingi-3 element with a *Z* score of 30. Significant *Z* values were also obtained when L1Tca was compared with the N terminus of the *pol* protein, upstream of the RT domain, from the *I* factor (*Z* = 10.4) (Abad *et al*., 1989), *Bombyx* (*Z* = 11.7) (Xiong & Eickbush, 1988) and Tad1-1 (*Z* = 7) (Cambareri *et al*., 1994). The highly conserved domains of the Ape family also have homology with the non-LTR elements described above (Figure 3).

L1Tcb (ORF2) shows high homologies with RT-related sequences of non-LTR retrotransposons

The FASTA and TFASTA programs used to search for similarities between L1Tcb and sequences present in the SWISSPROT and GENEMBL banks indicated that there are high homologies between L1Tcb and the RT domains of non-LTR retrotransposons. The BESTFIT program showed that the RT-related sequence from 12 out of the 15 non-LTR retrotransposons studied had a *Z* value greater than 10 when compared with L1Tcb and an identity of 21% to 28%. The highest homology was found with the *T. brucei* ORF2 from the ingi-3 retrotransposon with an identity of 28% and a *Z* value of 27. The trypanosomatid non-LTR retrotransposons CZAR, SLACS and CRE1 (Aksoy *et al*., 1990; Gabriel *et al*., 1990; Villanueva *et al*., 1991) showed *Z* values lower than 4. Interestingly, these elements are site-specific retrotransposons. *Z* values lower than 3 were detected between L1Tcb and the RT-related proteins of LTR-retrotransposons and retroviruses. After a PILEUP analysis using two RT from viruses, four RT-related proteins from non-LTR retrotransposons, and L1Tcb (Figure 4), we observed that L1Tcb maintains all seven motifs conserved in the RT and RT-related proteins (Toh *et al*., 1983; Hattori *et al*., 1986), and that, moreover, 36 of the 42 conserved identical or chemically similar amino acids described by Xiong & Eickbush (1990) are present in identical positions in L1Tcb. Among the most highly conserved residues are those that make up the ''Y / FXDD box'' typical of all RTs. The alanine at the X position detected in L1Tcb is characteristic of all non-LTR retrotransposons. The dendrogram derived from the PILEUP analysis of the RT domains from five LTR retrotransposons, four internal sequences belonging to group II introns, 14 non-LTR retrotransposons and L1Tcb indicates that L1Tcb fits into the non-LTR retrotransposon branch, *T. brucei* ingi-3 being the closest element.

L1Tcc (ORF3) contains two cysteine motifs

The analysis of the deduced amino acid sequence from ORF3 (L1Tcc) showed some of the characteristics of the *gag*-coded proteins, since two cysteine motifs and a high (7.5%) and non-uniform distribution of proline residues were found. Sixteen of the 21 proline residues were found flanking and inside the cysteine motifs. The BLASTP program revealed that the cysteine motifs of L1Tc with the $CX_2CX_{12}HX_{3-5}H$ structure were similar to the C_2H_2 class of zinc fingers from the transcription factors of high eukaryotic genomes (Pieler & Theunissen, 1993; O'Halloran, 1993). Interestingly, the same $CX_2CX_{12}HX_{3-5}H$ structure was also found in all the described trypanosomatid non-LTR retrotransposons, and in the insect R2Bm elements (Figure 5) instead of the $CX_2CX_4HX_4C$ (C_2HC class) in viruses and most retrotransposons.

Genomic organization and copy number of L1Tc

The chromosomal location of L1Tc-homologous sequences was resolved by (PFG) electrophoresis. The autoradiogram (Figure 6(a)), after 3 hours of running time, showed that the L1Tc sequences are dispersed among several size classes of chromosomes. The 0.9, 1.6 and 1.9 Mb long chromosomes, and the giant chromosomes are the most intensively labeled ones. After ten hours of running time, the labeling extended all over the chromosomal set. The genomic organization was revealed by hybridization of the DNA with the entire L1Tc element after digestion with different restriction enzymes. As expected from the chromosomal location, the autoradiogram (Figure 6(b)) showed that the

Figure 2

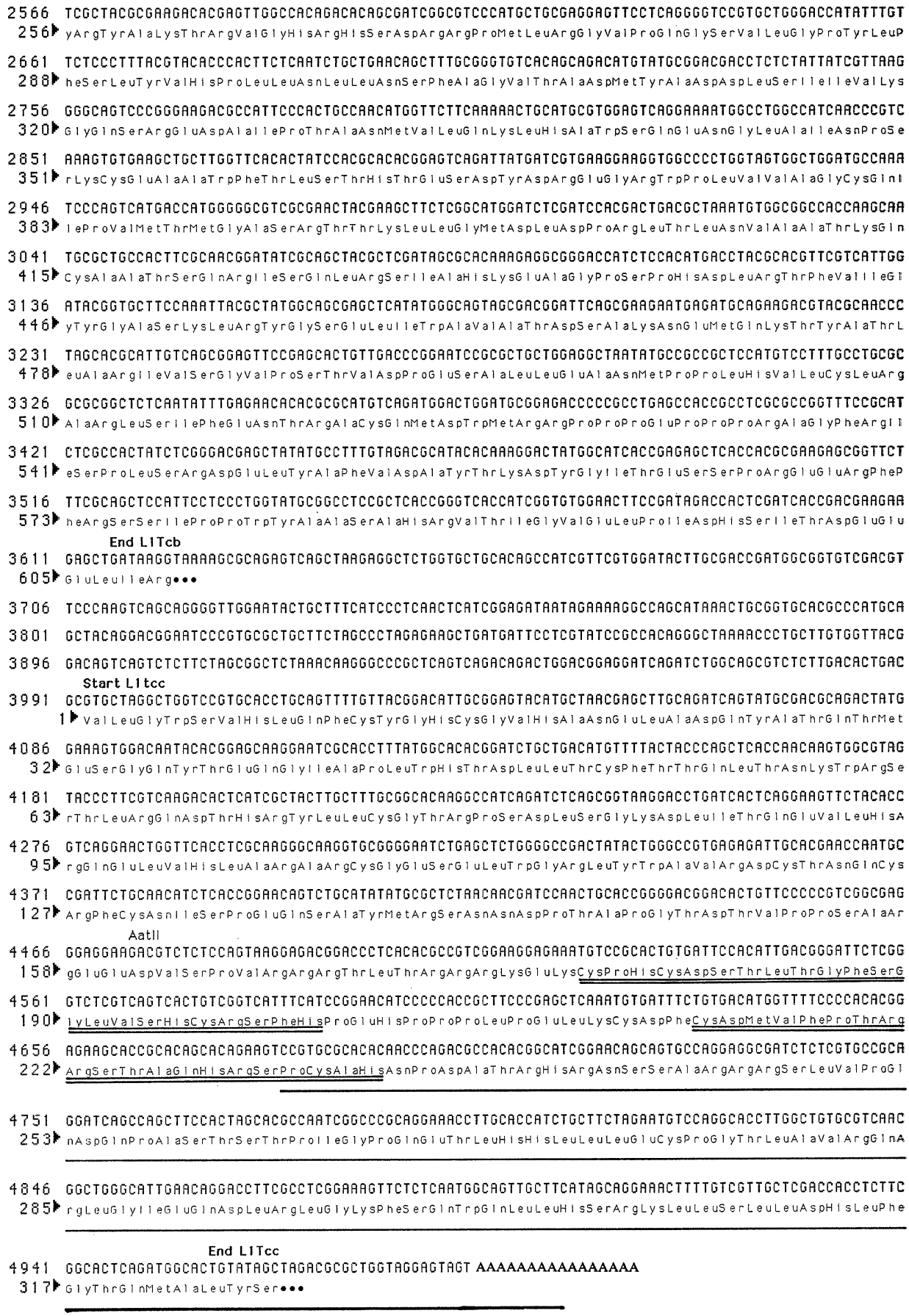

Figure 2. Complete nucleotide and deduced amino acid sequence of L1Tc. Each open reading frame (ORF) is translated below the DNA sequence, L1Tca (ORF1), L1Tcb (ORF2) and L1Tcc (ORF3). The stop termination codons are indicated by filled circles. The RIME-like sequence, showing homology with the RIME sequence of several transposons, at the 5' end and the E12A sequence at the 3' end are underlined. The *Eco*RI site at the 5' end and the *Aat*II at the 3' end are indicated. The 2 $CX_2CX_{12}HX_{3\text{-}5}H$ motifs are double underlined.

Figure 3. Comparison of L1Tc with the Ape family of proteins. The alignment of the proteins was performed using the PILEUP program. The numbers at the right- and left-hand sides of each sequence indicate the position of the amino acids in the proteins compared. The proteins and their sources are as follows: Ap1 is the major human apurinic endonuclease (Demple *et al*., 1991). Rrp1 is the 252 amino acid C-terminal region of the recombination repair protein 1 from *Drosophila* (Sanders *et al*., 1991). Exoa is the major DNA exonuclease of *Streptococcus pneumoniae* (Puyet *et al*., 1989). Exo3 is the exonuclease III from *Escherichia coli* (Saporito *et al*., 1988). Blocks of amino acid sequence identity (identical or chemically similar) are boxed. The homology between the *I* factor (Abad *et al*., 1989), ingi (Kimmel *et al*., 1987) and Tad1-1 (Cambareri *et al*., 1994) with the Ape family of proteins is also indicated. The amino acid sequence identity is indicated by asterisks.

L1Tc-related sequences are dispersed throughout the genome. A highly labeled band of about 5.5 kb was present in all lanes in which the DNA was digested with a restriction enzyme that cuts only once within L1Tc. The 5.5 kb long band hybridized with several probes from along the E12 element (Requena *et al*., 1994) as an indication that this band contains the entire E12. Analysis of chromosomal location and genomic organization of L1Tc in different strains of *T. cruzi* showed a great variability (unpublished results). The copy numbers of the L1Tc-homologous sequences were estimated by dot-blotting using the 3' *Aat*II-*Aat*II and the 5' *Eco*RI-*Eco*RI fragments of L1Tc as probes (Figure 6(c)). The pSPFM55 plasmid containing the entire element was used as reference. On the basis of the total genome content per parasite given by Borst (1982), the copy number was calculated to be about 2800 or 2300 depending on the probe used (3' or 5' probes, respectively). Therefore, it is likely that the number of 5' truncated elements would be about 17% of the total.

Discussion

In this paper, we have presented evidence showing that L1Tc is a non-LTR retrotransposon, and that it is actively transcribed into poly(A)$^+$ RNA. The L1Tc element is present in a high copy number and found dispersed throughout the genome of *T. cruzi*. This cDNA contains at its 3' end the E12A repetitive sequence (Requena *et al*., 1994), and at its 5' end a RIME-like sequence. Although RIME and RIME-like sequences are present in the VSG transcripts and the genomic ingi elements of *T. brucei*, in our knowledge this is the first report of a non-LTR retrotransposon

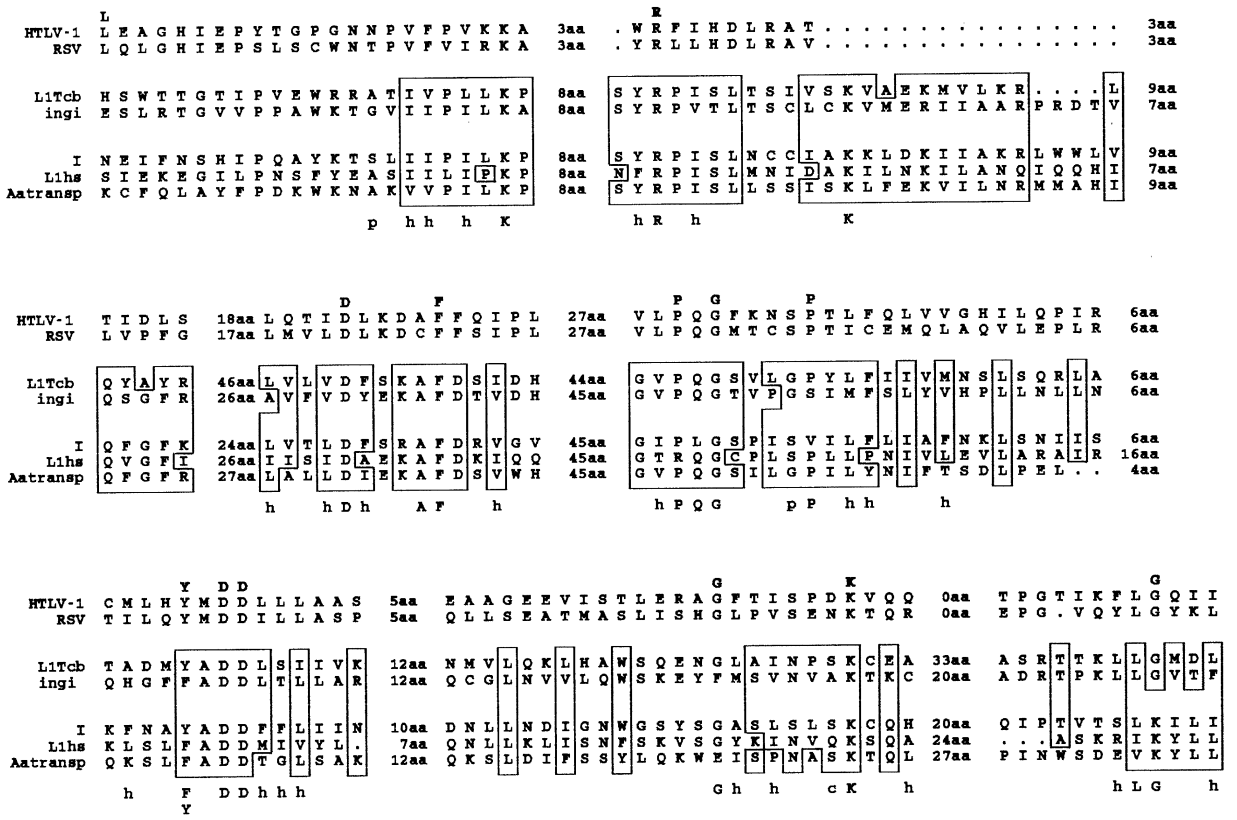

Figure 4. Comparison of L1Tcb with the amino acid (aa) sequence of the (RT) domain of 4 non-LTR retrotransposons and 2 retrovirus *pol* genes. The amino acid residues (in bold) at the top of the sequence correspond to invariable residues from a retrovirus (Toh *et al*., 1983; Hattori *et al*., 1986). The 42 conserved positions containing identical or chemically similar amino acids in all the RT sequences described by Xiong & Eickbush (1990) appear at the bottom. Boxed sequences are shown when the same amino acid residue appears in the same position in at least 4 of the 5 non-LTR retrotransposons. The numbers indicate the length of the gaps (in amino acids). The sequences used for the alignment were taken from: the 300 RT-related region of HTLV, the human T-cell leukaemia virus type 1 polymerase (Malik *et al*., 1988); RSVP, the Rous sarcoma virus polymerase (Schwartz *et al*., 1983; ingi, the ingi-3 non-LTR retrotransposon from *T. brucei* (Kimmel *et al*., 1987); *I*, the *I* factor from *Drosophila* (Abad *et al*., 1989); L1Hs, a LINE sequence from human (Xiong & Eickbush, 1990); Aatransp, the Juan-1 element, a LINE retroposon from *Aedes* mosquito species (Mouches *et al*., 1992).

cDNA that contains a RIME-like sequence. We think that there are no associated LTRs in L1Tc, since the tandemly repeated genomic 5.5 kb long band (Figure 6(b)), containing the 5.0 kb long cloned L1Tc cDNA, also contains the entire E12 element described by Requena *et al*. (1994). A 3' poly (A) stretch should be present in genomic L1Tc, since, as indicated by Requena *et al*. (1994), a poly (A) stretch is present at the 3' end of the genomic E12A element, which is located at the 3' of L1Tc. The existence of a highly

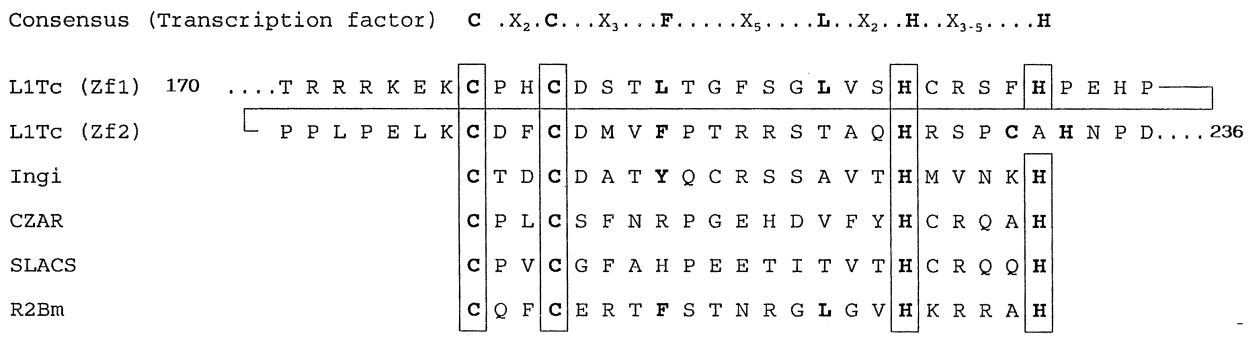

Figure 5. Comparison of the cysteine motifs found in the ORF3 of L1Tc (L1Tcc) with those found in ingi-3 (Kimmel *et al*., 1987), CZAR (Villanueva *et al*., 1991), SLACS (Aksoy *et al*., 1990) and R2Bm (Xiong & Eickbush, 1988). The consensus TFIIIA-like transcription factor motif (Pieler & Theunissen, 1993) is indicated above the Figure. The 2 cysteine motifs of L1Tcc, separated by 11 amino acids, are labeled as Zf1 and Zf2. The numbers indicate the position of the amino acids within L1Tcc. The Cys and His residues of the 4 non-LTR retrotransposons and L1Tcc are boxed.

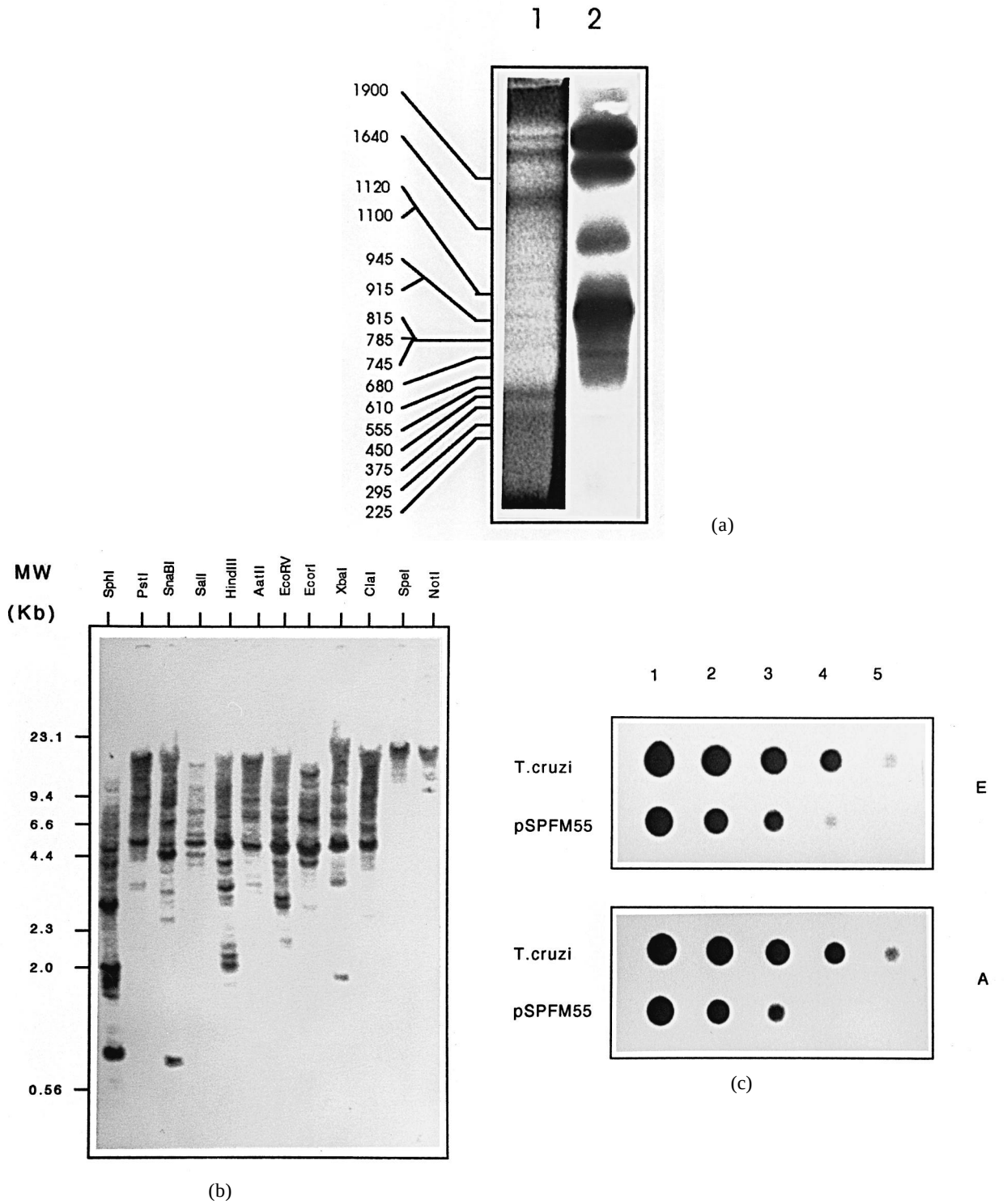

Figure 6. Genomic organization, distribution and copy number determination of the L1Tc element. (a) PFGE of *T. cruzi* chromosomes. Lane 1, ethidium bromide-stained gel; lane 2, Southern blot of the same gel hybridized with the L1Tc probe. The numbers at the left-hand side indicate the molecular mass markers (in Da) from *S. cerevisiae* chromosomes (kb). The autoradiogram was exposed for 3 h.

(b) Southern hybridization of the DNA from *T. cruzi* digested with several enzymes and probed with the L1Tc element. The *Pst*I, *Sal*I, *Hin*dIII, *Aat*II, *Eco*RV, *Eco*RI, *Xba*I and *Cla*I enzymes cut only once in the element. The numbers at the left-hand side indicate in kb the size of mobility of the *Hin*dIII fragments of lambda phage DNA.

(c) *T. cruzi* genomic content of the L1Tc element. Known amounts of *T. cruzi* DNA (3.33, 1.66, 0.83, 0.416 and 0.208 μg representing 10^7, 5×10^6, 2.5×10^6, 1.25×10^6 and 6.25×10^5 parasites, respectively) and pSPFM55 plasmid DNA (111, 55.5, 27.75, 13.87, and 6.94 ng representing 10^{10}, 5×10^9, 2.5×10^9, 1.25×10^9, and 6.25×10^8 copies of L1Tc, respectively) were dot-blotted onto a nylon membrane and hybridized with the L1Tc 3' end *Aat*II-*Aat*II probe (A) and with the L1Tc 5' end *Eco*RI-*Eco*RI probe (E).

conserved region within the *T. brucei* RIME sequence and the *T. cruzi* L1Tc RIME-like sequence, together with the recent demonstration of the existence of an internal promoter in several non-LTR elements (Eickbush, 1992), may suggest the possibility that the L1Tc RIME-like region could function as an internal promoter. In fact, it has been suggested that the insertion of a RIME may condition the activation of potential expression sites in *T. brucei* (Pays *et al*., 1989).

The most characteristic features shared by all the retrotransposons is a long ORF containing a RT-related sequence. All seven domains conserved in the RT and RT-related proteins (Toh *et al*., 1983; Hattori *et al*., 1986) are found in L1Tcb. The dendrogram derived from the PILEUP analysis of several groups of RT sequences (LTR, non-LTR and internal sequence to group II introns) indicates that L1Tcb fits into the non-LTR retrotransposon branch, the *T. brucei* ingi-3 element being the most closely related. Moreover, 36 of the 42 conserved amino acids described by Xiong & Eickbush (1990) to be present in RTs are found in identical positions in L1Tcb. This conservation corresponds to a value of 85.7%, which is similar to that reported by these authors for the non-LTR retrotransposon family. Among the most highly conserved residues are those that make up the specific ''Y/FXDD'' box of the RTs, which is necessary for their activity (Larder *et al*., 1987). The presence of alanine in the X position in L1Tcb reinforces the hypothesis that L1Tc is a non-LTR retrotransposon element, since the LTR retrotransposons and the viral RTs have a hydrophobic residue in the X position.

The presence of cysteine motifs in L1Tc also support the theory that the element is related to the non-LTR retrotransposon family. Most importantly, the cysteine motifs of L1Tc are located at a 3' position, as in the *T. brucei* ingi-3 element and some other non-LTR retrotransposons (Fawcett *et al*., 1986; Fanning & Singer, 1987; Schwarz-Sommer *et al*., 1987). In contrast, most of the retrotransposons contain the cysteine motifs in a 5' location. Since the cysteine motifs of L1Tcc, with the $CX_2CX_{12}HX_{3-5}H$ structure, show homology with the zinc fingers of some transcription factors (Pieler & Theunissen, 1993; O'Halloran, 1993), it is possible that there is a functional or evolutionary relationship between the cysteine motifs of L1Tcc and the zinc fingers of transcription factors. On the other hand, as it has been demonstrated that the $CX_2CX_{12}HX_{3-5}H$ structures are able to bind RNA (Pieler & Theunissen, 1993), besides their ability to bind DNA (O'Halloran, 1993), we suggest that the cysteine motifs present in L1Tc may be involved in functions similar to those of the *gag*-like protein.

We found that the ORF1 of the *T. cruzi* L1Tc element may encode for a polypeptide with endonucleolitic activity, since it shows high homology with the Ape family of repair enzymes implicated in the first step of repair of the AP sites (recognition of AP sites and DNA cleavage as hydrolytic AP endonucleases; Lindah, 1990). Thus, it is likely that L1Tca might have a similar function to that found for the enzymes of the Ape family, and that this could be a general feature of certain subsets of non-LTR retrotransposons. Interestingly, the most highly conserved domains in both L1Tca and the Ape family of proteins are also present upstream of the RT domains of *pol* genes from some non-LTR retrotransposons. Then, following the model for retrotransposition of non-LTR retrotransposons proposed by Eickbush (1992), the L1Tc as well as other non-site-specific non-LTR retrotransposons may have the capacity to generate DNA nicks in AP sites, where the integration complex could associate.

Methods

Trypanosomes

The Tulahuen strain of *T. cruzi* was used. Growth and maintenance of epimastigotes were as described (Nogueira *et al*., 1981).

DNA sequencing and analysis of the cDNA insert (L1Tc) in the pSPFM55 clone

The nucleotide sequencing of the DNA was carried out by the dideoxy chain termination method described by Sanger *et al*. (1977) using Sequenase (United States Biochemical). The complete sequence of the cDNA was obtained by subcloning and by the use of internal primers synthesized during the sequencing procedure. The FASTA and TFASTA programs (Lipman & Pearson, 1985) were used to search for similarities between the cDNA sequence and the sequences present in the GENEMBL and SWISSPROT data banks. Other programs from the GCG package and the BLAST network service of NCBI (Altschul *et al*., 1990) were also used for analysis and alignments. The statistical significance (*Z* score) was determined (Doolittle, 1981) after comparison of the sequence under investigation with 100 randomly permuted versions of the potentially related sequence. The PILEUP program of the GCG package was used to generate a dendrogram of several RT and RT-related proteins and the deduced amino acid sequence of the L1Tc.

Southern and Northern blot analysis

The *T. cruzi* DNA was isolated from epimastigotes with proteinase K and phenol/chloroform (1/1 v/v) and the RNA by the guanidinium thiocyanate method (Maniatis *et al*., 1989). After digestion with a variety of restriction enzymes the DNA fragments were separated in 0.8% agarose gels and transferred to nylon membranes (Zeta-probe, Bio-Rad). The RNA was size-separated in agarose/formaldehyde gels and transferred to nylon membranes. Hybridization for either the DNA or RNA analysis, was performed at 65°C overnight in 0.5 M NaH_2PO_4 (pH 7.2), 1 mM EDTA, 7% SDS and 0.25 mg/ml of herring sperm DNA. The probes were labeled by the random primed method (Feinberg & Vogelstein, 1983). Washing of the filters was done in 40 mM NaH_2PO_4 (pH 7.2), 1 mM EDTA, 5% SDS twice for 30 minutes at 65°C and twice in 40 mM NaH_2PO_4 (pH 7.2), 1 mM EDTA, 1% SDS for the same period of time and at the same temperature. For reprobing, the membranes were washed twice for 20 minutes at 95°C in 500 ml of 0.1 × SCC (0.15 M NaCl, 0.015 M sodium citrate (pH 7.0), 0.5% SDS.

Copy number determination

Different amounts of total *T. cruzi* genomic DNA and the pSPFM55 plasmid DNA (constructed in the pSPORT1 vector; Requena *et al*., 1994) were denatured in 0.2 M NaOH, 0.2 M EDTA at 90°C for five minutes. After denaturation, the DNA samples were loaded by duplicate on Immobilon-N membranes (Millipore) using a Millipore Dot Blot apparatus. The filters were hybridized with the probes (*Eco*RI-*Eco*RI and *Aat*II-*Aat*II) under the conditions described above. As a control, the probes were also hybridized with the vector. The copy number was calculated by densitometric analysis of the autoradiogram using the pSPFM55 plasmid DNA as reference, and on the basis of the nuclear genome content per parasite (Borst *et al*., 1982) after subtraction of the signal given by the control.

Pulsed field gradient electrophoresis

Agarose blocks containing approximately a total of 5×10^7 parasites were prepared (Clark *et al*., 1990) and stored in 0.5 M EDTA (pH 9.5; 1/5 block was subjected to electrophoresis (1% agarose in 0.5× TBE at 200 V for 24 hours at 12°C with a pulse time of 120 seconds. The DNA was transferred to nylon filters and hybridized with the L1Tc probe as described above.

Acknowledgements

We thank Philip Buchers from the ISREC (Lausanne) for computer assistance. This work was supported by BIO90-0786 and BIO93-0043 grants from CICYT Plan Nacional I + D, Spain. The sequence reported in this paper has been deposited in the EMBL database (accession no. X83098).

References

Abad, P., Vaury, C., Pélissoon, A., Chaboissier, M. C., Busseau, I. & Bucheton, A. (1989). A long interspersed repetitive element—the *I* factor of *Drosophila teissieri*—is able to transpose in different *Drosophila* species. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8887–8891.

Affolter, M., Rindisbacher, L. & Braun, R. (1989). The tubulin gene cluster of *Trypanosoma brucei* starts with an intact β-gene and ends with a truncated β-gene interrupted by a retrotransposon-like sequence. *Gene,* **80**, 177–183.

Aksoy, S., Williams, S., Chang, S. & Richards, F. F. (1990). SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs. *Nucl. Acids Res.* **18**, 785–792.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bontempi, E. J., Búa, J., Aslund, L., Porcel, B., Segura, E. L., Henriksson, J., Orn, A., Petterson, U. & Ruiz, A. M. (1993). Isolation and characterization of a gene from *Trypanosoma cruzi* encoding a 46-kilodalton protein with homology to human and rat tyrosine aminotransferase. *Mol. Biochem. Parasitol.* **59**, 253–262.

Borst, P., Van der Ploeg, L. H. T., Van Hock, J. F. M., Tas, J. & James, J. (1982). On the DNA content and ploidy of Trypanosomes. *Mol. Biochem. Parasitol.* **6**, 13–23.

Buschiazo, A., Campetella, O. E., Macina, R. A., Salceda, S., Frasch, A. C. & Sanchez, D. O. (1992). Sequence of the gene for a *Trypanosoma cruzi* protein antigenic during the chronic phase of human Chagas disease. *Mol. Biochem. Parasitol.* **54**, 125–128.

Cambareri, E. B., Helber, J. & Kinsey, J. A. (1994). Tad1-1, an active LINE-like element of *Neurospora crassa. Mol. Gen. Genet.* **242**, 658–665.

Clark, C. G., Lai, E. Y., Fulton, C. & Cross, G. A. M. (1990). Electrophoretic karyotype and linkage groups of the amoeboflagellate *Naeglaria gruberi. J. Protozool.* **37**, 400–408.

Demple, B., Herman, T. & Chen, S. D. (1991). Cloning and expression of APE, the cDNA encoding the major human apurinic endonucelase: definition of a family of DNA repair enzymes. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 11450–11454.

Doolittle, R. (1981). Similar amino acid sequences: chance or common ancestry. *Science*, **214**, 149–159.

Eickbush, T. H. (1992). Transposing without ends: The non-LTR retrotransposable elements. *New Biol.* **4**, 430–440.

Fanning, T. & Singer, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucl. Acids Res.* **15**, 2251–2260.

Fawcett, D. H., Lister, E. K., Kellet, E. & Finnegan, D. J. (1986). Transposable elements controlling I–R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. *Cell,* **47**, 1007–1015.

Feinberg, A. P. & Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132**, 6–13.

Gabriel, A., Yen, T. J., Schwartz, D. C., Smith, C. L., Boeke, J. D., Sollner-Webb, B. & Cleveland, D. W. (1990). A rapidly rearranging retrotransposon within the miniexon gene locus of *Crithidia fasciculata. Mol. Cell. Biol.* **10**, 615–624.

González, A., Prediger, E., Huecas, M. E., Nogueira, N. & Lizardi, P. M. (1984). Minichromosomal repetitive DNA in *Trypanosoma cruzi*: its use in a high-sensitive parasite detection assay. *Proc. Nat. Acad. Sci. U.S.A.* **81**, 3356–3360.

Hasan, G., Turner, M. J. & Cordingley, J. S. (1982). Ribosomal RNA genes of *Trypanosoma brucei*. Cloning of a rRNA gene containing a mobile element. *Nucl. Acids Res.* **10**, 6747–6760.

Hasan, G., Turner, M. J. & Cordingley,, J. S. (1984). Complete nucleotide sequence of an unusual mobile from *Trypanosoma brucei. Cell,* **37**, 333–341.

Hattori, M., Kuhara, S., Takenaka, O. & Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature (London),* **321**, 625–628.

Hobbs, M. R. & Boothroyd, J. C. (1990). An expression-site-associated gene family of trypanosomes is expressed in vivo and shows homology to a VSG gene. *Mol. Biochem. Parasitol.* **43**, 1–16.

Hutchinson, C. A., Hardies, S. C., Loeb, D. D., Shehee, W. R. & Edgell, W. H. (1989). in *Mobile DNA* (Berg, D. E. & Howe, M. M., eds), pp. 593–617, Amer. Soc. Microbiol., Washington, DC, U.S.A.

Kimmel, B. E., Onesmo K., Ole, M. & Young, J. R. (1987). Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. *Mol. Cell. Biol.* **7**, 1465–1475.

Larder, B. A., Purifoy, D. J. M., Powell, K. L. & Dambi, G. (1987). Site-specific mutagenesis of AIDS virus reverse transcriptase. *Nature (London)* **327**, 716–717.

Leeton, P. R. J. & Smyth, D. R. (1993). An abundant

LINE-like element amplified in the genome of *Lilium speciosum*. *Mol. Gen. Genet.* **237**, 97–104.

Lindahl, T. (1990). Repair of intrinsic DNA lesion. *Mutat. Res.* **238**, 305–311.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Malik, K. T. A., Even, J. & Karpas, A. (1988). Molecular cloning and complete nucleotide sequence of an adult T cell leukaemia virus / human T cell leukaemia virus type I (ATLV / HTLV-I) isolate of Caribbean origin: relationship to other members of the ATLV / HTLV-I subgroup. *J. Gen. Virol.* **69**, 1695–1710.

Maniatis, T., Fritsch, E. F. & Sambroock, J. (1989). In *Molecular Cloning. A Laboratory Manual*. 2nd edition, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., U.S.

Morse, B., Rotherg, P. G., South, V. J., Spandorfer, J. M. & Astrin, S. M. (1988). Insertional mutagenesis of the3 *myc* locus by a LINE-1 sequence in a human breast carcinoma. *Nature (London)*, **333**, 87–90.

Mouches, C., Bensaadi, N. & Salvado, J. C. (1992). Characterization of a LINE retrosposon dispersed in the genome of three non-sibling Aedes mosquito species. *Gene*, **120**, 183–190.

Murphy, N. B., Pays, A., Tebabi, P., Coquelet, H., Guyaux, M., Steinert, M. & Pays, E. (1987). *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties. *J. Mol. Biol.* **195**, 855–871.

Nogueira, N., Chaplan, S., Tydings, J. D., Unkeless, J. & Cohn, Z. (1981). *Trypanosoma cruzi*. Surface antigens of blood and culture forms. *J. Exp. Med.* **153**, 629–639.

O'Halloran, T. V. (1993). Transition metals in control of gene expression. *Science*, **261**, 715–725.

Pays, E., Tababi, P., Pays, A., Coquelet, H., Revelard, P., Salmon, D. & Steinert, M. (1989). The genes and transcripts of an antigen gene expression site from *T. brucei*. *Cell*, **57**, 835–845.

Pieler, T. & Theunissen, O. (1993). TFIIIA: nine fingers–three hands?. *Trends Biochem. Sci.* **18**, 226–230.

Puyet, A., Greenberg, B. & Lacks, S. A. (1989). The *exoA* gene of *Streptococcus pneumoniae* and its products, a DNA exonuclease with apurinic endonuclease activity. *J. Bacteriol.* **171**, 2278–2286.

Requena, J. M., Jimenez-Ruiz, A., Soto, M., Lopez, M. C. & Alonson, C. (1992). Characterization of a highly repeated interspersed DNA sequence of *Trypanosoma cruzi*: its potential use in diagnosis and strain classification. *Mol. Biochem. Parasitol.* **51**, 271–280.

Requena, J. M., Martín F., Soto, M., López, M. C. & Alonso, C. (1994). Characterization of a short interspersed reiterated DNA sequence of *Trypanosoma cruzi* located at the 3'-end of a poly(A)$^+$ transcript. *Gene*, **146**, 245–250.

Sanders, M., Lowenhaupt, K. & Rich, A. (1991). *Drosophila* Rrp1 protein: an apurinic endonuclease with homologous recombination activities. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 6780–6784.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci. U.S.A.*, **74**, 5463–5467.

Saporito, S., Smith-White, B. J. & Cunninghan, R. P. (1988). Nucleotide sequence of the *xth* gene of *Escherichia coli* K-12. *J. Bacteriol.* **170**, 4542–4547.

Schwartz, D., Tizard, R. & Gilbert W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell*, 853–869.

Schwarz-Sommer, Z., Leclercq, E. G. & Saedler, H. (1987). Cin4, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. *EMBO J.* **6**, 3873–3880.

Singer, M. F. (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell*, **28**, 433–434.

Singer, M. F. & Skowronski, J. (1985). Making sense out of LINEs: long interspersed sequences in mammalian genomes. *Trends Biochem. Sci.* **10**, 119–122.

Toh, H., Hayashida, H. & Miyota, T. (1983). Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. *Nature (London)*, **305**, 827–829.

Villanueva, M., Williams, S. P., Beard, C. B., Richards, F. F. & Aksoy, S. (1991). A new member of a family of site-specific retrotransposons is present in the splice leader RNA genes of *Trypanosoma cruzi*. *Mol. Cell. Biol.* **11**, 6139–6148.

Weiner, A. M., Deininger, P. L. & Efistratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55**, 631–661.

Xiong, Y. & Eickbush, T. H. (1988). The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non–long–terminal–repeat retrotransposons. *Mol. Cell. Biol.* **8**, 114–123.

Xiong, Y. & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362.

Edited by K. Yamamoto

(Received 8 June 1994; accepted 12 December 1994)