REVIEW ARTICLE

Mobile genetic elements in protozoan parasites

SUDHA BHATTACHARYA[1]*, ABHIJEET BAKRE[1] and ALOK BHATTACHARYA[2]

[1]*School of Environmental Sciences and* [2]*School of Life Sciences, Jawaharlal Nehru University,*
New Delhi 110 067, India

Abstract

Mobile genetic elements, by virtue of their ability to move to new chromosomal locations, are considered important in shaping the evolutionary course of the genome. They are widespread in the biological kingdom. Among the protozoan parasites several types of transposable elements are encountered. The largest variety is seen in the trypanosomatids—*Trypanosoma brucei*, *Trypanosoma cruzi* and *Crithidia fasciculata*. They contain elements that insert site-specifically in the spliced-leader RNA genes, and others that are dispersed in a variety of genomic locations. *Giardia lamblia* contains three families of transposable elements. Two of these are subtelomeric in location while one is chromosome-internal. *Entamoeba histolytica* has an abundant retrotransposon dispersed in the genome. Nucleotide sequence analysis of all the elements shows that they are all retrotransposons, and, with the exception of one class of elements in *T. cruzi*, all of them are non-long-terminal-repeat retrotransposons. Although most copies have accumulated mutations, they can potentially encode reverse transcriptase, endonuclease and nucleic-acid-binding activities. Functionally and phylogenetically they do not belong to a single lineage, showing that retrotransposons were acquired early in the evolution of protozoan parasites. Many of the potentially autonomous elements that encode their own transposition functions have nonautonomous counterparts that probably utilize the functions in *trans*. In this respect these elements are similar to the mammalian LINEs and SINEs (long and short interspersed DNA elements), showing a common theme in the evolution of retrotransposons. So far there is no report of a DNA transposon in any protozoan parasite. The genome projects that are under way for most of these organisms will help understand the evolution and possible function of these genetic elements.

[Bhattacharya S., Bakre A. and Bhattacharya A. 2002 Mobile genetic elements in protozoan parasites. *J. Genet*. **81**, 73–86]

Introduction

Long before biologists were afforded the precision of genome sequencing, it was known that certain DNA sequences in the genome are mobile. These are collectively called transposable elements or transposons. By virtue of their ability to move to new locations, transposons are considered to be potent agents in shaping the evolutionary course of the genome. Transposable elements are usually divided into two major groups according to their mechanism of transposition (Finnegan 1989; Kazazian 1998). The first group is composed of DNA-based transposable elements that transpose via a DNA intermediate (figure 1). These include the P elements of *Drosophila melanogaster*, the Ac and Spm elements of maize, and the Tc1 transposon of *Caenorhabditis elegans*. The second group, the retroelements, move by reverse transcription of an RNA intermediate. DNA transposons are common in bacteria, invertebrates and plants, whereas in the vertebrates retrotransposons are more abundant. Retrotransposons fall into two major classes—the long terminal repeat (LTR)-containing and the non-LTR-containing elements (Boeke and Corces 1989). The LTR elements have remarkable structural and functional resemblance to retroviruses, and are exemplified by the Ty-1 and Ty-2 elements of *Saccharomyces cerevisiae* and Copia of *D. melanogaster*. Compared with the LTR elements, the non-LTR elements use a fundamentally different mechanism for transposition, called target-primed reverse transcription (Luan *et al*. 1993; Finnegan 1997) (figure 2). Prominent members of this class are the mammalian long interspersed elements

*For correspondence. E-mail: sb@mail.jnu.ac.in.

Keywords. protozoan parasites; autonomous and nonautonomous mobile elements; non-LTR retrotransposons.

(LINEs) and short interspersed elements (SINEs), the *D. melanogaster* I and Jockey elements and the *Bombyx mori* rDNA elements. Active copies of transposons that encode *trans*-acting functions required for transposition are relatively rare since they destabilize the genome by insertional mutagenesis. Most copies accumulate multiple mutations, which render them inactive in terms of gene expression. Autonomous elements that encode their own transposition functions also aid the transposition of smaller, nonautonomous elements that have acquired the necessary *cis* specificities to utilize the proteins encoded by the autonomous elements.

The variety of transposons encountered in nature is truly impressive. They occupy as much as 50% of genome content in some organisms; yet their function, if any, remains largely speculative (Kazazian 2000). They are very widely distributed throughout the biological kingdom. Here we concern ourselves with the group of organisms called the protozoan parasites. The advent of large-scale genome sequencing has been particularly beneficial to the study of protozoan parasites where conventional genetic tools are ill developed. A common feature of parasite genomes, which is not observed in most free-living organisms, is the unusual plasticity of the genome. This is manifested in chromosome-length polymorphisms such that homologous chromosomes may vary in size by as much as 50–100%, and in rare cases even by 400%. The size variation is brought about by recombination and breakage–repair mechanisms, especially in the repetitive sequences comprising the subtelomeric regions of chromosomes. However, genome sequencing shows that some of the polymorphism may also be contributed by the spread of mobile genetic elements.

In this article we briefly review the available information on transposons in protozoan parasites (summarized in table 1). The trypanosomes emerge as clear winners by being the most extensively studied parasites in this regard. Somewhat surprisingly, not a single report is yet

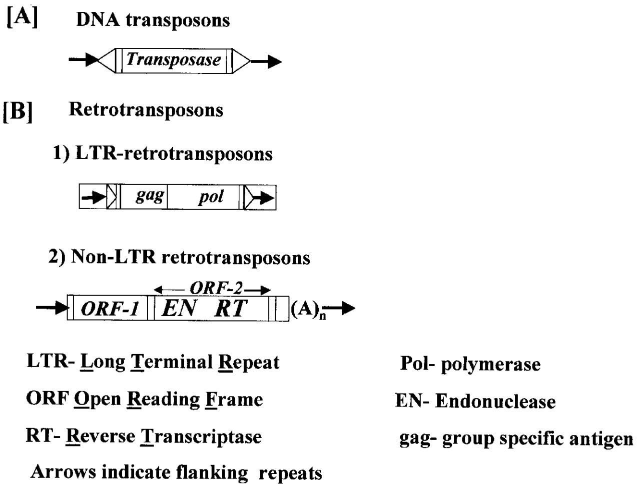

Figure 1. Two classes of mobile genetic elements found in genomes.

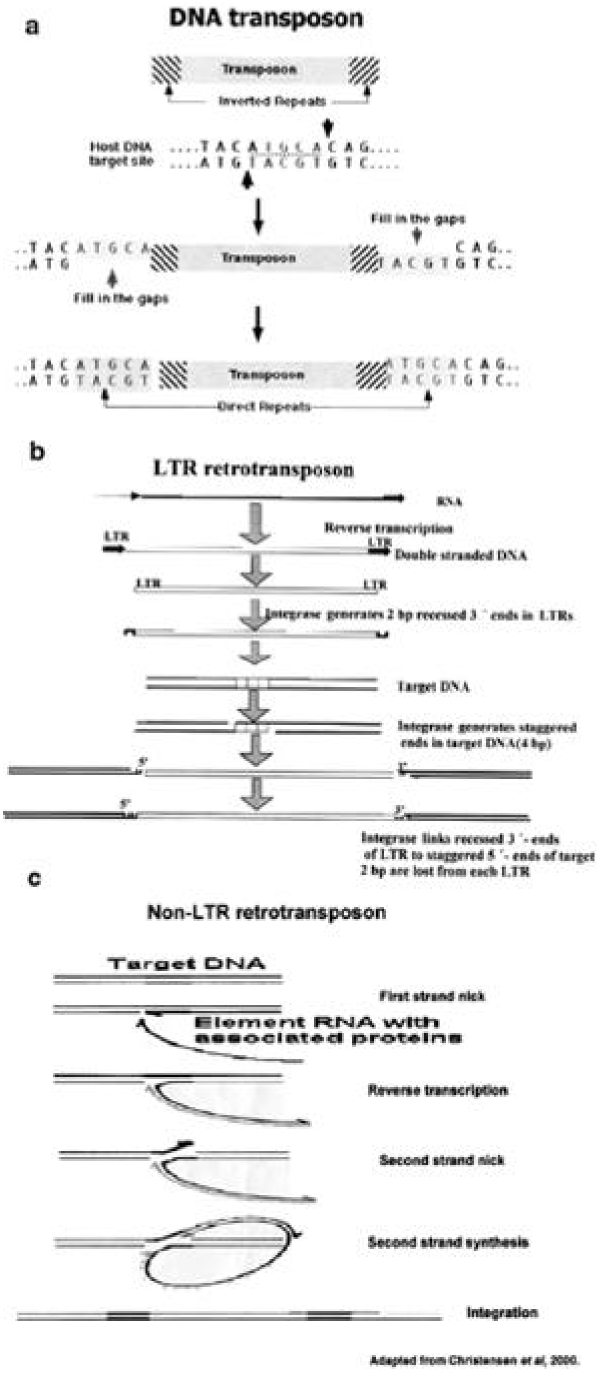

Figure 2. Mechanism of transposition for (a) DNA transposons, (b) LTR retrotransposons, (c) non-LTR retrotransposons. (Figure 2c used with permission from Christensen S., Pont-Kingdon G. and Carroll D. 2000 Target specificity of the endonuclease from the *Xenopus laevis* non-long terminal retrotransposon, Tx1L. *Mol. Cell. Biol.* **20**, 1219–1226. © American Society for Microbiology.)

Table 1. Salient features of retrotransposons in protozoan parasites.

	T. brucei			*T. cruzi*				*C. fasciculata*		*G. lamblia*		*E. histolytica*	
	INGI	RIME	SLACS	L1Tc	CZAR	SIRE	Viper	CRE1	CRE2	GilM & GilT	GilD	EhRLE1	IE / Ehapt2
Location	Near repeatitive genes	Near repetitive genes	SL-RNA genes	Dispersed	SL-RNA genes	Dispersed; frequently near telomeres	Four chromosomes (0.96 to 1.6 Mb)	SL-RNA genes	SL-RNA genes	Subtelomeric	Near repetitive genes	Dispersed	Dispersed
Copy no.	~ 500	~ 500	9	2800	30–40	1500–3000	~ 300	10	6	~ 15	30	~ 140	~ 500
Size (kb)	5.2	0.512	6.678	5.5	7.237	0.428	2.539	3.5	9.6	~ 6.0	~ 3.0	4.804	0.55
TSD (bp)*	12	12	49		22			29	29	None		Variable size	Variable size
Transcript (kb)	2.5 to 9.0	0.8 to 8.0	Not detected	5 (some small species)	Not detected	0.8 to 7.0	Not detected	Not detected				1.9 (Full size not detected)	0.6
5′-UTR (bp)	Short, ~ 100		1511	101	1504		707	None	844	55	None	14	
3′-UTR (bp)			148	535	534			60	1200	3000	Short	17	
No. of ORFs	One	One	Two	Three	Two		One	One	One	One	One	Two	
Domains in ORFs													
RT* location, motif	Central, FADD		ORF2, YLDD	ORF2, YADD	ORF2, YLDD			YIDD	Central, YIDD	Central		ORF2, YMDD	
EN* location, motif	N-terminus, apurinic endonuclease			ORF1, apurinic endonuclease	ORF1, CCHH				C-terminus, CCHC	C-terminus, CCHC REL-ENDO		ORF2 C-terminus, CCHC match with restriction endonucleases	
NB* location, motif	C-terminus, CCHH		ORF1 C-terminus, CCHH	ORF3, CCHH	ORF1, CCHH; ORF2 N-terminus, HHCC				N-terminus, CCHH; C-terminus, CCHH	N-terminus, CCHH, 2 motifs	N-terminus, CCHH	ORF1	
Other							RNaseH						
Element type	Non-LTR	Non-LTR	Non-LTR	Non-LTR	Non-LTR	LTR	LTR	Non-LTR	Non-LTR	Non-LTR	Non-LTR	Non-LTR	Non-LTR
Potentially autonomous?	Yes	No; may need INGI	Yes	Yes	Yes	No; may need VIPER	Yes	Yes (?); may need CRE2	Yes	Yes	All present-day copies inactive	Yes	No; may need EhRLE1

*Abbreviations: TSD, target site duplication; UTR, untranslated region; RT, reverse transcriptase; EN, endonuclease; NB, nucleotide-binding domain.

documented on the existence of transposons in *Plasmodium* and *Leishmania*, two of the best-studied protozoans that infect humans. It would be too early to interpret this as a lack of mobile genetic elements in these organisms. Transposons do exist in other protozoan parasites like *Giardia lamblia*, *Crithidia fasciculata* and *Entamoeba histolytica*; and in each they exhibit unique properties.

Trypanosoma brucei

The African trypanosomes, including *Trypanosoma brucei*, are responsible for diseases affecting humans (African sleeping sickness) and livestock. The organism is diploid, with a haploid genome size of about 35 megabases (Mb) (El-Sayed *et al*. 2000). Chromosomes can be grouped in three size classes—(i) 11 pairs of megabase chromosomes ranging in size from 1 Mb to 6 Mb, (ii) several intermediate chromosomes of 200 kilobases (kb) to 900 kb and uncertain ploidy, and (iii) about 100 minichromosomes of 50–150 kb. The megabase chromosomes contain the protein-coding genes, and the telomere-linked expression sites for variable surface glycoprotein (VSG) genes. The intermediate chromosomes and minichromosomes also serve as repositories for telomere-linked VSGs. These genes constitute 5% or more of the genome. The minichromosomes are composed primarily of tandem arrays of a 177-bp repeat, which constitutes 10–20% of the genome. About 5% of the *T. brucei* genome consists of a retrotransposon, the INGI/RIME element.

INGI / RIME

INGI (which stands for ‘many’ in Kiswahili) is a 5.2-kb non-LTR retrotransposon (Kimmel *et al*. 1987). It consists of a 4.7-kb central region that is flanked at the two ends by 250-bp sequences that are actually the two halves of a 512-bp transposable element called RIME (ribosomal mobile element) (Hasan *et al*. 1984). INGI/RIME display two features of a non-LTR retrotransposon, namely presence of an oligo-dA tail at the 3′-end and insertion site duplication. At about 500 copies per haploid genome, INGI/RIME are the most abundant putatively mobile elements in the *T. brucei* genome. These elements are located on megabase chromosomes but not on intermediate chromosomes and minichromosomes. Their distribution shows that they are associated with a number of highly repeated gene families, including rRNA genes, tandemly repeated tubulin gene arrays, and RHS (retrotransposon hot spot), a large, multigene family (128 copies per haploid genome) encoding mainly nuclear proteins with an ATP/ GTP binding motif (Bringaud *et al*. 2002). About a third of RHS pseudogenes contain RIME or INGI or both inserted in frame at the same nucleotide position. The genome sequence of chromosome Ia of *T. brucei* shows that INGIs/RIMEs precede or are within most expression sites for the VSG genes. These sites are close to the telomeres (El-Sayed *et al*. 2000). Studies with RHS genes show that the elements are inserted preferentially at a 12-mer sequence (TACTGTTATACA), which is repeated at both ends following insertion (Bringaud *et al*. 2002). This creates new insertion hot spots for the elements and results in their tandem insertion. The tandem arrangement of RIME / INGI (and the *T. cruzi* L1Tc element described later) is unique, since none of the site-specific or randomly distributed retroelements in other systems show this organization.

Of the INGIs sequenced so far none have been found to be identical with one another. All have accumulated mutations that are scattered throughout the element. It is not established whether either INGI or RIME gave rise to the other, and a mechanism by which this may have occurred has not been elucidated. There may be several possibilities. The central 4.7-kb region may have got deleted from INGI to make the first RIME. The deletion could have occurred at the level of splicing from an RNA transcript, or at the DNA level. However, there are no splicing signals in INGI at the appropriate points; nor are there any direct repeats that might have mediated DNA recombination. Alternatively, INGI could have been derived by insertion of another mobile element into RIME. No remnant of a target site duplication that may result from such an event has been found. Whatever be the mechanism, RIME retains the two ends of INGI which may contain the necessary *cis* specificities for transposition, but depends on the enzymatic activity encoded by INGI; much like the autonomous mammalian LINEs are thought to aid the transposition of their nonautonomous counterparts, the SINEs (Weiner 2000).

INGI encodes a single large protein containing a central reverse transcriptase (RT) domain, a C-terminal DNA binding domain and an N-terminal apurinic (AP) endonuclease domain. The potential size of the translated open reading frame (ORF) is 1646 amino acids. Sequence alignment shows good match with RT of retroviruses and retrotransposons. The conserved YXDD (amino acid single-letter code) box in the RT active site is FADD in this element (Murphy *et al*. 1987). Comparative analysis of deduced amino acid sequence of the INGI RT domain shows that INGI is more closely related to mammalian LINEs than to other retrotransposons of trypanosomes, e.g. CZAR, SLACS and CRE (Lodes *et al*. 1993).

Northern blot analysis of INGI transcripts showed a heterogeneous population of molecules ranging in size from 2.5 to 9 kb (Vassella *et al*. 1996). No discrete 5.2-kb transcript was seen. RIME-specific probes also gave a series of bands from 0.8 kb to ~ 8.0 kb. Antisense probes also yielded faint signals. Transcription was moderately a-amanitin sensitive. Transcripts were much more abundant in the form of the parasite that exists in the mammalian bloodstream than in the procyclic form which is in the insect vector. Transcripts of mammalian L1 are also heterogeneous in size, but only a minor population of uniform-length transcripts are transported to the cytoplasm. In contrast, most RIME and INGI specific heterogeneous RNA is in the cytoplasm. Analysis of the 5′-ends of 19 cDNA clones of these elements showed that some of these are probably derived from read-through transcription from far-upstream promoters. A second group of cDNAs, which start with a short conserved sequence, may be derived from transcripts that use internal promoters. Internal promoters are thought to be used in other non-LTR elements as well, e.g. the *D. melanogaster* Jockey element (Mizrokhi *et al*. 1988). A 39-nucleotide miniexon or spliced leader sequence is added to the 5′-ends of all nuclear-encoded mRNAs of trypanosomes by a process of *trans*-splicing. However, there was no *trans*-splicing of miniexon to INGI/RIME transcripts. *Trans*-splicing in trypanosomes is controlled by pyrimidine-rich sequences upstream of the 3′ splice site. The complete absence of *trans*-splicing in RIME/INGI transcripts shows that these elements probably never transpose in the vicinity of a splicing signal. Translation of the transcripts may be mediated by an internal ribosome entry site.

In general, transposons promote a variety of large-scale recombination events including translocations, duplications and deletions. Since INGI/RIME are inserted very frequently in the vicinity of VSG genes, which are responsible for antigenic variation, a possible role of these elements may be in the evolution of VSG gene repertoires by promoting recombination (Kimmel *et al*. 1987).

SLACS

As stated above, a common feature of mRNA structure in all trypanosomatids is the posttranscriptional addition to all pre-mRNAs of a common 39-nucleotide sequence by *trans*-splicing. This spliced leader, or miniexon sequence, is at the 5′-end of a small nonpolyadenylated transcript, the spliced leader (SL) RNA (Agabian 1990). SL-RNA genes exist in 200–600 copies in tandem arrays in discrete genomic loci. A universal feature of SL-RNA genes in trypanosomatids seems to be the existence of interrupting sequences that insert in a subset of these genes (Aksoy *et al*. 1987).

SLACS (spliced leader-associated conserved sequence) is a non-LTR retrotransposon that is present in only nine copies, all of which are inserted between nucleotides 11 and 12 from the 5′-end of the SL-RNA genes of *T. brucei* (Aksoy *et al*. 1990). The element is 6678 bp long, with an extensive polyA stretch at the 3′-end. It generates a 49-bp target site duplication upon insertion. Two ORFs span 75% of the sequence. ORF1 encodes 384 amino acids (nucleotides 1512–2726), and ORF2 encodes 1182 amino acids (nucleotides 2802–6530). In the region upstream of ORF1 (nucleotides 462–1173) there are three tandem repeats of 185 bp. The sequence between nucleotides 1180 and 1512 has no stop codon. Therefore ORF1 could be longer by 110 amino acids; however, the first methionine codon is at position 1512.

ORF1 polypeptide, like the gag polypeptide of retroviruses, may be a nucleic-acid-binding protein. It contains a variation of the CCHH motif associated with the metal-binding domain found in many DNA-binding proteins. This motif $CX_2CX_{13}HX_5H$ (where X is any amino acid), is also encoded by the INGI ORF. In the latter the corresponding nucleotide sequence is repeated five times in the last third of the ORF. In SLACS it is present once at the 3′-end of ORF1.

ORF2 encodes a protein with an RT domain. In a comparative analysis of RT domains from various retrotransposons, Xiong and Eickbush (1988) marked eight regions in the RT domain in which the non-LTR retrotransposons show greater similiarity to each other than to retroviruses or Copia-like elements. Of 32 invariant residues in these regions, 24 are identical in SLACS. Two residues in region four contain chemically similar amino acids. Five residues in region five have conservative amino acid substitutions. In the YXDD box, SLACS encodes L for X while all other non-LTRs encode A. Leucine (L) is encoded in this position in the RT of LTR elements.

A unique feature of SLACS is that all nine copies are conserved in sequence. There are no truncated copies. The nine copies differ in the number of 185-bp repeats and the number of A residues at the 3′-end. However, all have the same 49-bp target site duplication at the two ends. The differences show that the nine copies may not have arisen from a single copy by unequal sister chromatid exchange.

The 185-bp repeats at the 5′-end of SLACS may serve as promoters. In the mouse non-LTR retrotransposon L1, there are 200-bp repeats located at the 5′-end of ORF1. These contain promoter activity (shown by expression experiments with gene fusion constructs) and vary in number in different L1s. In SLACS, promoter activity has not yet been demonstrated experimentally in these repeats. In addition, no SLACS-specific RNA transcript could be detected. It is possible that the transcript may be of low abundance or may be unstable, or it may be expressed only in certain stages of the life cycle.

Trypanosoma cruzi

This parasite is the agent of Chagas' disease, a chronic, incapacitating disease prevalent in most of Latin America. The haploid genome size is about 40 Mb, with a GC content of 48–50%. Definition of molecular karyotype has been hampered by the extensive chromosomal-size polymorphism. The CL Brener strain contains 12 megabase chromosomes ranging from 3.5 to 1.0 Mb and eight intermediate-sized ones ranging from 1.0 to 0.45 Mb

(Cano *et al*. 1995). A number of non-LTR retrotransposons have been described in this organism. The L1Tc element of *T. cruzi* is similar to INGI/RIME of *T. brucei* while the counterpart of SLACS in *T. cruzi* is CZAR.

L1Tc

Named after a *T. cruzi* cDNA, L1Tc is a 5.5-kb element, present in 2800 copies per genome, and accounts for 17% of the total genome. It is dispersed throughout the genome (Martin *et al*. 1995). It is actively transcribed into polyA+ RNA. The element itself has a stretch of As at its 3′-end. At the 5′-end it has a RIME-like sequence of 127 nucleotides, which is 69% identical with the RIME of *T. brucei*. Since RIME insertion is thought to activate potential expression sites, an internal promoter may exist at the 5′-end of L1Tc. Nucleotide sequence analysis of L1Tc showed that the element has three ORFs in three reading frames. ORF1 (frame 1) spans nucleotides 102 to 1228, ORF2 (frame 2) nucleotides 1799 to 3623, and ORF3 (frame 3) nucleotides 3993 to 4965.

ORF1 polypeptide has homology with the APE family of nucleases involved in DNA repair. It also showed similarities with the corresponding domains in *T. brucei* INGI polypeptide and the N-terminus of proteins encoded by some other non-LTR retrotransposons.

ORF2 has high homology with RT domains of non-LTR elements. The highest homology was found with *T. brucei* INGI, followed by SLACS, CZAR and CRE-1 (described later). There was much less homology with RT of LTR retrotransposons and retroviruses. Thirtysix out of 42 conserved identical or chemically similar amino acids in RT described by Xiong and Eickbush (1990) are present in identical positions in ORF2 polypeptide. The YXDD box typical of all RTs is present; X is alanine in case of L1Tc. Alanine is most commonly encoded at this position in non-LTR retrotransposons. The YADD motif lends further support to the view that L1Tc is a non-LTR type of retroelement.

ORF3 polypeptide has features of gag-like protein of retroviruses. It has two cysteine motifs ($CX_2CX_{12}HX_{35}H$) similar to the CCHH class of zinc fingers seen in transcription factors of higher eukaryotes. A similar motif is also seen in the other trypanosomatid retroelements (INGI, CZAR and SLACS) and in the insect R2Bm elements (Xiong and Eickbush 1988), while a CCHC motif is encoded in viruses and retrotransposons. In addition, the cysteine motif code is located at the 3′-end of the element (as in other non-LTR elements) in contrast to the 5′-end location seen in other types of elements. Proteins with cysteine motifs (like gag) are known to have RNA-binding properties. ORF3 polypeptide may be involved in RNA binding through this motif.

The ORF1 polypeptide of L1Tc has been functionally characterized in some detail. The gene has been cloned and expressed in *Escherichia coli* as a 40-kDa protein (Olivares *et al*. 1997). It contains three conserved domains with similiarity higher than 65% to the same domains of APE consensus sequence. The recombinant 40-kDa protein had demonstrable APE activity, and the cloned gene could complement *E. coli* mutants lacking exoIII activity (an AP enzyme). This protein may play a role in generating free 3′-OH sites in chromosomal DNA, which may constitute the first stage of the transposition of the element. A high number of potential AP sites can be generated along chromosomal DNA, which may be acted upon by this enzyme. This may explain the high copy number and dispersion of L1Tc in the genome. In addition to APE activity, the L1Tc ORF1 also has 3′-phosphatase and 3′-phosphodiesterase activities, which effectively remove 3′-blocking groups from damaged DNA substrates (Olivares *et al*. 1999). Given the potential involvement of element-encoded nucleases in integration of the elements themselves, the 3′-phosphatase and phosphodiesterase may allow 3′-blocking ends to function as targets for insertion of L1Tc, in addition to AP sites. On the other hand, the 3′-repair activities may actually indicate a possible repair role for L1Tc. In fact, repair of double-stranded DNA breaks because of insertion of Ty-1 element from *S. cerevisiae* has been described (Teng *et al*. 1996).

Phylogenetic analysis of the ORF1 sequence showed no correlation with the corresponding sequences from human L1 elements. L1Tc was much closer to INGI and was in a branch different from that of the APE family. In general the non-site-specific non-LTR elements are distributed randomly along the tree, reinforcing the hypothesis that these elements have either spread horizontally among major taxonomic groups, or that their origin predates the evolution of metazoans (Xiong and Eickbush 1990).

CZAR

CZAR (cruzi-associated retrotransposon) is a 7237-bp element (Villanueva *et al*. 1991). Like SLACS of *T. brucei*, it inserts specifically into SL-RNA genes between nucleotides 11 and 12 of the SL 39-mer. The insertion is accompanied by a 22-nucleotide direct repeat (positions 11–32 of the SL gene) at the site of insertion. Nucleotides 14–30 of the SL gene are repeated at positions 4–20 of the element. Southern hybridization of chromosomes separated by pulsed field gradient electrophoresis (PFGE) showed that CZAR was located on a single chromosome of 1200 to 1300 kb on which all SL-RNA genes reside. There are 300 copies of SL-RNA genes and 30–40 copies of CZAR per diploid genome. Thus, 10% of all SL-RNA genes have a CZAR insertion. The 5′-UTR of the element (1504 bp) contains a 185-nucleotide repeat present in two full and one partial copies. The repeat number varies in different copies, leading to size polymorphism. A similar repeat arrangement is also seen in the 5′-UTR of SLACS. The 3′-UTR of CZAR is a 534-bp stretch followed by 42 A residues. Two additional insertion sites of CZAR were studied in which different copies of the element were inserted. They also had the same 22-bp-long target site duplications of identical sequence. This shows that either the insertion events in the various copies of SL genes are recent, or that functional constraints are being exerted to maintain the target site duplications. This contrasts with other non-LTR elements in which the extent and composition of duplications varies between different copies (Xiong and Eickbush 1988; Hutchison *et al*. 1989).

The element contains two potential ORFs. ORF1 (nucleotides 1505–2663) encodes 386 amino acids, and ORF2 (2741–6692) encodes 1317 amino acids. ORF1 polypeptide of CZAR, like its counterpart in SLACS, has a CCHH motif in the middle ($CX_2CX_{12}HX_4H$) with nucleic-acid-binding properties. It is different from the motif in gag domains and is very similar to the motif seen in *Xenopus* transcription factor TF IIIA (Miller *et al*. 1985). INGI encodes the same motif but at the 3′-end of its single ORF.

ORF2 polypeptide contains an RT domain in its central region. According to the scheme of Xiong and Eickbush (1988, 1990), 23 of 35 invariant residues are conserved and the spacing between the eight conserved regions of RT is also the same. The RTs of CZAR, SLACS and CRE1 are remarkably similar, with 55% amino acid identity among them. The RTs of CZAR and SLACS show 75% identity. In the YXDD box, X is leucine in CZAR, in which respect this element is similar to retroviruses. The similiarity between CZAR and SLACS extends further downstream of the RT domain, where CRE1 diverges from the other two. Towards the N-terminus of ORF2 polypeptide is a DNA-binding domain with an HHCC motif ($HX_4HX_{19–20}CX_2C$). This is encoded in various retroviruses and in some LTR elements and is thought to be characteristic of endonuclease/integrase domains (Johnson *et al*. 1986). It is also found in SLACS and CRE1 products. Apart from this, CZAR, SLACS and CRE1 products do not show any other similiarity with retroviral endonuclease domains. Such an endonuclease motif is not found in other non-LTR elements (e.g. R1, R2, I, F and L1Md).

The overall organization of CZAR is strikingly similar to that of SLACS and CRE1, particularly to SLACS. All three elements are located in discrete chromosomal loci in low copy numbers. They insert between nucleotides 11 and 12 of the SL 39-mer. Nucleotides 1–12 of the 39-mer are identical in all three. All contain unusually long target site duplications (22 in CZAR, 49 in SLACS, 28 in CRE1) compared with 4–14 in other non-LTR elements. CZAR and SLACS both contain 185-nucleotide repeats in the 5′-UTR. Similar repeats are also found in mouse L1Md and are thought to represent internal promoters (Loeb *et al*. 1986). In CZAR and SLACS, ORF1 is separated from ORF2 by 78 and 79 nucleotides respectively, and both ORFs are in the same frame. As in the case of SLACS, no CZAR-specific transcripts could be found in *T. cruzi*. The current data are insufficient to draw any significant conclusions in this regard.

SIRE / VIPER

Unlike L1Tc and CZAR, for which closely related elements were found in *T. brucei*, the SIRE/VIPER class of transposons is *T. cruzi*-specific with no homologues found in other *Trypanosoma* species (Vazquez *et al*. 1994, 2000).

SIRE (short interspersed repeat element) is present in about 1500–3000 copies per genome, distributed on all chromosomes (Vazquez *et al*. 1999). It is frequently located near telomeres (Chiurillo *et al*. 1999) and is also linked to protein-coding genes. The element is 428 bp in length. The central portion of the element is most conserved and has high GC content. The 3′-end of the element contains a functional *trans*-splicing signal. About 2.2% of mRNAs screened from a cDNA library contain SIRE. The size of these mRNAs ranges from 800 to 7000 bp. No 428-nucleotide transcript was seen in Northern blots, indicating that the element is not transcribed as such. SIRE is most often located at the 3′-end of mRNAs, on the sense strand of transcription. In 63% of insertions, SIRE provides the polyadenylation site, while in the remaining 37% the polyadenylation site is located 50 bp downstream of SIRE insertion. When SIRE provides the polyadenylation site, it is located in a hot spot at the 3′-end of the element. When located immediately upstream of a protein-coding gene, SIRE can donate an SL-acceptor site to the gene, since it has a functional SL-acceptor sequence at its 3′-end (Vazquez *et al*. 1994).

SIRE is related to VIPER (vestigial interposed retroelement), a 2359-bp element. The 3′-end of VIPER contains a part of the SIRE sequence (220 bp from middle to 3′-end of SIRE) and the 5′-end contains 182 nucleotides from 5′-end to middle of SIRE. Therefore VIPER contains the whole of SIRE (in two parts, like RIME is contained in INGI), except that 30 bp from the middle of SIRE are missing. Approximately 300 copies of VIPER are present in the genome, distributed on four chromosomal bands of 0.96 to 1.6 Mb. The reconstructed ORF of VIPER starts at position 708 and can code for a protein of 487 amino acids. The protein contains RT and RH (RNAaseH) motifs separated by a 103-residue tether. From alignment of RT sequences, the element is close to the LTR retrotransposon DROME of *D. melanogaster* (Tchurikov *et al*. 1989) and to the rice tungro bacilliform virus (Hay *et al*. 1991). It does not share homology with non-LTR RTs. However, LTRs that flank the coding region of LTR retrotransposons have not been identified in VIPER. SIRE may be a nonautonomous element of which the autonomous counterpart is VIPER, which encodes RT and RH functions that can be utilized in *trans*. In this context it is interesting that the 3′-ends of SIRE and VIPER have the same sequence, which may be required in *cis* for transposition.

An analysis of 150 SIREs shows high sequence homology. This conservation implies that SIREs are successful genomic 'parasites', or that positive selection has imposed genomic maintenance. The functional properties of SIRE show that the latter may be true. For example SIRE can donate SL-acceptor site to downstream genes and polyadenylation site to upstream genes. These features of SIRE are important because in *T. cruzi* the polyadenylation site and surrounding intergenic sequences play a role in stage-specific gene expression. SIRE and VIPER may also promote recombination events leading to chromosome translocation, duplication and deletion. A duplication event with the arrangement SIRE–gene–SIRE–gene–SIRE has been identified (Vazquez *et al*. 1994). SIREs are also found at 3′-ends of genes in subtelomeric regions where differentially expressed genes for surface antigens are known to reside (Chiurillo *et al*. 1999). The translocation of a ribosomal pseudogene to a different chromosome was also associated with VIPER in the vicinity of the pseudogene (Vazquez *et al*. 1999). Therefore these elements may be involved in gene shuffling. The fact that they are found only in *T. cruzi* may be relevant to the evolution of this parasite.

Crithidia fasciculata

A member of the Trypanosomatidae family, it is an insect parasite. Being noninfectious to humans it is widely studied as a model organism.

As in *T. brucei* and *T. cruzi*, translatable mRNAs in *C. fasciculata* possess an identical 39-nucleotide leader sequence at their 5′-termini called the miniexon or spliced leader sequence (Gabriel *et al*. 1987). The miniexon sequence is also found at the 5′-end of an abundant transcript called the miniexon donor RNA. In *C. fasciculata* this RNA is 90 nucleotides long and is transcribed from a family of multicopy tandemly arrayed miniexon genes which have a unit length of 423 bp and a copy number of 200 to 500 per genome.

CRE1

This element (*Crithidia fasciculata* retrotransposable element 1) is 3.5 kb long and inserts specifically in the miniexon gene locus, within the 39-bp miniexon sequence (Gabriel *et al*. 1990). The insertion leads to a target site duplication of 29 nucleotides. Multiple copies of CRE1, in the same orientation, are interspersed among the tandemly repeated miniexon genes. Some CRE1 copies also exist tandemly, uninterrupted by miniexon genes. There are 10 copies of CRE1 per genome. The insertion pattern of CRE1 in the miniexon gene array is very dynamic. When cells were grown from individual colonies and CRE1 patterns tested from 60 colonies by Southern hybridization, no two colonies had the same pattern. PFGE analysis was also done on cells grown from individual colonies. *C. fasciculata* has at least 12 chromosomes ranging in size from 450 kb to > 1.2 Mb. CRE1 was located on chromosomes in the size range 600–900 kb and the pattern varied in different clones. All of the chromosomes that contained CRE1 also contained miniexon genes but the converse was not true. By using appropriate restriction enzymes it was clear that not only are CRE1 and miniexon genes on the same chromosomes but they are also physically linked. Subclones were obtained from a clonal population after growth for 30 generations. CRE1 patterns were seen by Southern hybridization of genomic DNA digested with appropriate restriction enzymes. About 30% of subclones had CRE1 patterns distinct from the parental pattern. The change included concomitant appearance and disappearance of one or more bands. The rate of rearrangement was estimated to be at least 1% per generation. Therefore CRE1 is a rapidly rearranging element.

CRE1 has a single ORF of 3420 nucleotides followed by a string of As. The length of the polyA tract at the 3′-end of the element varies from 16 to 57 nucleotides. The ORF begins one nucleotide downstream of the 5′-end of the element and ends 20 nucleotides upstream of the polyA tract. It could potentially be transcribed in continuity with the upstream miniexon gene. The first AUG is at codon number 380 of the ORF. In five independently cloned elements, the first 80 codons were sequenced and no AUG was found. Unless all of them carry the same mutation, this would mean that a different initiation codon is used for translation. The CRE1 ORF encodes a highly conserved RT domain with a YXDD box (X is isoleucine). This region encodes a functional RT when cloned and expressed as a fusion product with the yeast retrotransposon Ty-1 (Gabriel and Boeke 1991). Two potential nucleic-acid-binding domains could be identified upstream of the RT domain. No discrete RNA species corresponding to the element could be seen by Northern hybridization. CRE1 may be transcribed at very low levels, or its RNA may be unstable, or its expression may be regulated in an unknown way.

From sequence analysis the element resembles non-LTR retrotransposons. The target sites for integration of CRE1, SLACS, CZAR and LINS1 (a retrotransposon in *Leptomonas seymouri*) (Bellofatto *et al*. 1988) in the miniexon gene cluster are all overlapping, occurring within two bases of one another (between nucleotides 11 and 12) in the highly conserved 39 nucleotides encoding the miniexon sequence. This remarkable similarity in the target site used for insertion suggests that the miniexon locus in trypanosomatids may serve as a sink for reverse-transcribed genes. The conserved integration site may be the target of a site-specific endonuclease. Since CRE1 has a complete ORF, it may be a functional element capable of active transposition. It occurs in other *Crithidia* species as well, showing that it is not a recent intruder in the *C. fasciculata* genome. The rapid rearrangements of CRE1 revealed by clonal analysis may be a result of reciprocal or nonreciprocal recombination (gene conversion), or both, along with active transposition to other sites.

CRE2

Another element, related to CRE1 but clearly distinct from it, is found in the miniexon gene array of *C. fasciculata*. This element, called CRE2, is 9.6 kb long and inserts at precisely the same location as CRE1 in a subset of miniexon genes (Teng *et al*. 1995). As with CRE1, CRE2 insertion is also accompanied by 29-bp target site duplications, beginning at nucleotide 11 of the miniexon sequence. Like CRE1, CRE2 also shows genomic instability. There are six copies per genome, which are restricted to chromosomes that contain the miniexon arrays. CRE1 and CRE2 predominate on different size classes of chromosomes. The two elements may have evolved in separate host strains and may have been brought into the same genome, more recently, via a mating event.

A single ORF, predicted to encode a protein of 2518 amino acids, occupies 79% of CRE2. CRE1 and CRE2 have approximately 30% identity over a 1000-amino-acid region towards the C-terminus of the ORF. Beyond this the two elements are structurally distinct. Whereas CRE2 has a 844-bp 5′-UTR, CRE1 has no apparent 5′-UTR. CRE2 has a 1200-bp 3′-UTR while CRE1 has only a 60-bp 3′-UTR. As opposed to CRE1, CRE2 lacks the variable-length 3′-terminal polyA tracts. Therefore two evolutionarily diverged transposons share the same insertion site within the same genome. Although CRE2 3′-UTR does not end with a polyA tract, the UTR has internal poly(dA) tracts (nucleotides 9115–9132 and 9333–9344). The 5′-UTR also has a T tract (43 Ts) and an A tract (24 As). The ORF has the same orientation as the direction of transcription of the miniexon unit.

Database search shows that the highest similarity of CRE2 ORF is with CRE1, SLACS and CZAR. The C-terminal 1000 residues are the best conserved with an overall amino acid identity of 30%. This region includes two potential metal-binding motifs, CCHH and CCHC, in the presumptive endonuclease domain (codons 1517–1583), and a large domain (codons 1865–2116) corresponding to the eight segments of RT. In the YXDD domain of RT, the X is leucine or isoleucine in the products of these four elements, whereas in products from most other non-LTR elements it is alanine. Apart from the C-terminus of the encoded polypeptides, the structure of CRE2 is more similar to that of SLACS and CZAR than to CRE1. Although SLACS and CZAR both include two ORFs, while CRE2 has a single ORF, their respective coding capacities are comparable. The N-terminus of CRE2 product and the first ORF polypeptides of CZAR and SLACS have two regions of similarity. There is a proline-rich region, which may be involved in protein–protein interactions to form an aggregate or core particle. The second is a putative metal-binding domain (CCHH) at codon 405. This may be a DNA-binding domain as seen in the gag protein of retroviruses. Since CRE1 lacks the equivalent of the amino terminus of CRE2, it is an intriguing possibility that CRE2 provides structural, gag-like proteins in *trans* to CRE1, which might be important for replication of CRE1.

Like CRE1 and CRE2, two completely distinct families of site-specific retrotransposons R1 and R2 are inserted at different conserved sites in a fraction of the rRNA genes in most insects. Here too, the two elements have independently diverged in their individual sequences, while maintaining the target site specificity (Burke *et al*. 1993).There is evidence for other CRE-like elements also in *C. fasciculata*. Further analysis will reveal their relationships.

Giardia lamblia

It infects the small intestine of humans and a variety of other mammalian hosts (Adam 1991). *Giardia* species are notable for their lack of a number of organelles, including mitochondria, peroxisomes and nucleoli. A Golgi apparatus is not discernible in the trophozoite stage but can be seen during encystation. The genome size is about 12 Mb, with a GC content of 46%. Each trophozoite has two functionally equivalent nuclei. There are five chromosomes and each is present in at least four and perhaps eight or more copies. The telomeric regions are active in recombination, due to which chromosomes show multiple size variations.

Three families of non-LTR retrotransposons have been reported in *Giardia lamblia* (Arkhipova and Morrison 2001). Two of these, GilM and GilT, are potentially active elements while the third, GilD, is probably composed entirely of inactive copies.

GilM and GilT

Each family is represented mostly by intact elements, which show homology exceeding 99%, indicative of recent retrotransposition activity. Both families are confined to immediate subtelomeric regions. They are organized in tandem arrays in a head to tail orientation. Individual members in each array are separated from each other by stretches of $(A)_n$ (n = 10–16), with no target site duplication. The most distal member in the array is truncated at its 5′-end and capped by telomeric repeats $(TACCC)_n$.

The coding region of GilT and GilM consists of a single ORF that encodes about 1000 amino acids, with 54% identity and 67% similarity between the two families. The 5′-UTR is 55 bp. The ORF consists of a central RT domain, which contains nine conserved motifs found in other members of the LINE superfamily. The N-terminus of the polypeptide contains two zinc finger motifs of the CCHH type and the C-terminus contains a CCHC finger followed by the so-called REL-ENDO domain identified in CRE, SLACS and CZAR, and in insect R2 ribosomal insertion elements in which the site-specific insertion of the element is attributed to this domain.

About half the copies of each family carry a frameshift mutation between the CCHH and RT domains. Such frameshifts are common in other retroelements also and are thought to reduce the synthesis of RT (which may be necessary to modulate transposition activity). Such a frameshift mutation is also found in the HeT-A telomere-associated retrotransposon in *Drosophila* (Pardue *et al*. 1996).

The 3′-UTRs of GilM and GilT are unusually long, about 3 kb in length. The UTRs of the two elements have no sequence similarity except the 110-bp segment preceding the polyadenylation signal AGTAAA and the $(A)_n$ stretch. The 3′-UTR of GilM has a 750-bp sequence that is sometimes present as a tandem repeat. This segment contains 160 bp originating from the 3′-end of the large subunit of the ribosomal DNA in an antisense orientation. Read-through transcription and subsequent retrotransposition is a known property of human LINEs and may have contributed to 3′-UTR formation in *G. lamblia*.

An analysis of sequences adjoining *G. lamblia* telomeres showed that eight out of 11 sequences analysed were 5′-truncated copies of GilM and GilT. Two were rDNA sequences and one was a variant-specific surface protein. Most GilM arrays are followed at their 3′-end by single-copy genes, while some are adjacent to rDNA. GilT arrays are mostly followed by rDNA. The tandem arrangement of telomeric transposons may serve to protect proximal copies of genes from terminal degradation following telomere loss.

GilD

It is a high-copy-number family (about 30 per genome) in which most members have suffered extensive mutation. The reconstructed consensus ORF occupies most of the 3-kb element and has overall 25% identity and 39% similarity with GilM and GilT. The N-terminus of the encoded polypeptide is truncated and lacks one of the CCHH fingers. The 3′-UTR is short. The divergence of GilD sequences from the consensus is 6–13%, indicating inactivation in the distant past. Mutations are distributed throughout, without any bias towards synonymous sites. GilD is often located near variant-specific surface protein genes, pseudogenes or other repetitive genes such as ankyrins.

Phylogeny

Using RT for phylogeny, the *Giardia* retrotransposons appear to be closest to those containing the REL-ENDO domain, such as the NeSL and R2 clades (Malik and Eickbush 2000). GilD occupies a more basal position than GilM and GilT, and all three form a distinct clade, indicating that they were established in *Giardia lamblia* a long time ago.

Transposons at telomeres

Telomeric and subtelomeric regions are particularly suitable for transposon insertion because insertion events would not interfere with the functioning of nuclear genes. Instead, transposon insertions at these sites might be beneficial by expanding the buffer zone between the end of the chromosome and nearby genes. Transposons specific for telomeres/subtelomeres have been reported in *S. cerevisiae* (Ty-5), *Bombyx mori* (SART, TRAS), *Chlorella vulgaris* (Zepp) and *Allium cepa* (MIPT), besides *D. melanogaster* (HetA and TART). In *D. melanogaster* the retrotransposons have completely taken over the function of telomeres (Levskaya *et al*. 1994). There are no telomeric repeats and no telomerase in this organism. Healing of chromosome breaks occurs by terminal transposition of HetA and TART. Unlike *D. melanogaster*, *G. lamblia* has distinct telomeres, with GilM and GilT being subtelomeric.

Overall structural comparisons show that GilT and GilM are very similar to HetA and TART (Pardue *et al*. 2001). Both have the ability to form tandem arrays at chromosome ends and they have the coding capacity for RT and nucleic-acid-binding proteins. Both have an unusually long 3′-UTR in which the 3′-most region is prone to tandem duplication. The UTR may be needed for telomeric chromatin structure or terminal transposition activity or both. The long UTR in these two otherwise distant species (*G. lamblia* and *D. melanogaster*) may have appeared owing to convergent evolution towards some shared functionality (Pardue *et al*. 2001).

Entamoeba histolytica

It is a human pathogen, residing in the large intestine, and causes amoebiasis. Like *Giardia*, it lacks well-defined organelles typical of higher eukaryotic cells, e.g. mitochondria and Golgi bodies (Lushbaugh and Miller 1988). A closely related species *Entamoeba dispar* also colonizes the human large intestine but is not associated with disease (Diamond and Clark 1993). The haploid genome size is estimated to be about 20 Mb, with a ploidy of at least four. Chromosomes range in size from

0.3 to 2.2 Mb (Bagchi *et al*. 1999; Willhoeft and Tannich 1999) and have so far been resolved into 14 linkage groups (Willhoeft and Tannich 1999). Because of extensive chromosome length polymorphism, PFGE does not permit accurate resolution of chromosome number.

EhRLE1

In a study of a repetitive DNA sequence in *E. histolytica*, the first retrotransposon-like element (EhRLE1) was identified in this organism (Sharma *et al*. 2001). It is 4804 bp in size and is present in about 140 copies per genome. All copies show sequence variation with respect to one another (2–4% from the consensus). A complete ORF is missing in most copies. The complete EhRLE1 unit was reconstructed by sequence comparison. It potentially includes two nonoverlapping ORFs in two different reading frames. The 5′-UTR and 3′-UTR are short (14 and 17 nucleotides respectively).

The ORF1 polypeptide of EhRLE1 is 498 amino acids in length. ORF1 polypeptides of most non-LTR retroelements have a cysteine–histidine motif (CCHC, and less commonly CCHH). An exception is the polypeptide from the L1 element of mammals (Malik *et al*. 1999). EhRLE1 ORF1 polypeptide also does not have this motif, but it has homology with proteins with nucleic-acid-binding properties, and those mediating protein–protein interactions. This is similar to L1 ORF1 polypeptide, which has nucleic acid chaperone activity (Martin and Bushman 2001). ORF1 polypeptide is thought to bind RNA and form RNP particles.

ORF2 polypeptide (944 amino acids) has a central RT region, which is highly conserved; only three amino acids out of 42 in the seven RT domains deviate from the conserved positions in this region. The YXDD box in domain V, which is part of the RT active site, is present as YMDD. From RT sequence comparison, EhRLE1 appears to belong to the R4 clade of non-LTR retrotransposons, with its closest relatives being the R4 element of *Ascaris lumbricoides* and the Dong element of *Bombyx mori*. At its C-terminus the ORF2 polypeptide contains an endonuclease domain. It consists of a conserved CCHC motif ($CX_2CX_6HX_4C$) and an extensive region of similarity with members of the R2, R4 and CRE clades of non-LTR elements (Bagchi 2001). Of the 11 clades of non-LTR elements, these three differ from the remaining eight clades, in which the endonuclease domain belongs to the AP family of endonucleases (Malik *et al*. 1999). The EhRLE1 endonuclease domain (as in R2, R4 and CRE) has striking match with a conserved motif in a variety of restriction endonucleases. The endonucleases to which EhRLE1 has closest resemblance are all site-specific enzymes.

The R2 and R4 elements insert at a specific site in the large subunit gene of rRNA and the CRE elements insert in the spliced leader sequences. However, there is no evidence that EhRLE1 inserts in a site-specific manner in the *E. histolytica* genome. From PFGE analysis, it appears to have transposed to all the chromosomes of *E. histolytica* with no evidence of site specificity. The now-emerging genome sequence of *E. histolytica* shows that this element inserts in the vicinity of protein-coding genes and there is no evidence of its preferential location at telomeres. Considering the fact that EhRLE1 is otherwise very similar to the R2, R4 and CRE clades, its lack of insertion site specificity is a unique feature.

Although EhRLE1 does not integrate in a site-specific manner, it does have a high affinity for AT-rich sequences. Thus, although it may not recognize a specific sequence it may be specific for a particular DNA conformation. This is similar to human L1 endonuclease which integrates in a site-non-specific manner but specifically cleaves DNA with certain structural and sequence parameters, with minor groove width being of particular importance (Cost and Boeke 1998). The DNA sequence that best fits these requirements is a run of T_nA_n. Therefore EhRLE1 endonuclease seems to be functionally closer to the enzyme encoded by human L1.

EdRLE

EhRLE-like sequences could not be found in other *Entamoeba* species, including *E. moshkovskii*. However, the closely related, sibling species *E. dispar* does contain sequences that cross-hybridize with EhRLE. A 3-kb clone was obtained from an *E. dispar* genomic library and sequence analysis showed that it had a truncated copy of an element with very close homology to EhRLE1. It has been named EdRLE (Sharma *et al*. 2001). Its overall nucleotide sequence identity with EhRLE is about 85%. It has a conserved RT domain. Further data are awaited before the properties of this element can be understood.

IE/Ehapt2

A short repetitive sequence of 550 bp, named interspersed element (IE) or *E. histolytica* abundant polyadenylated transcript 2 (Ehapt2), has been reported by two groups (Cruz-Reyes *et al*. 1995; Willhoeft *et al*. 1999). It appears frequently in *E. histolytica* cDNA libraries, but all copies sequenced so far lack an ORF. The copy number of this element is estimated to be about 500 per genome. The sequence identity between different copies is of the order of 95%. A polyA+ transcript corresponding to the element is detected in Northern blots. The element is widely distributed in the genome and is frequently seen close to protein-coding genes. Its high copy number and wide genomic distribution make it a likely candidate for a mobile genetic element. Strong support for this view comes from the observations that the ends of this element are flanked by short direct repeats of the target site (U. Willhoeft and E. Tannich, unpublished) and that a stretch of 74 nucleotides at the 3′-end of this element is almost identical in sequence with the 3′-end of EhRLE1 (Bagchi 2001). This latter feature is seen in the SINE elements, which are nonautonomous and are thought to use the enzymatic machinery of LINEs for their transposition. The 3′-ends of SINEs are homologous to the 3′-ends of their respective LINEs (Boeke 1997; Okada *et al*. 1997). It is likely that the IE/Ehapt2 element and EhRLE constitute a LINE/SINE pair.

General considerations

Transposon silencing by RNAi

Double-stranded RNA can trigger the degradation of homologous RNA by a phenomenon called RNA interference (RNAi). Double-stranded RNAs are processed to produce 21–25-nucleotide small interfering RNAs (siRNAs), which target RNA transcripts for degradation. An intriguing observation has recently been made in *T. brucei*, where 24–26-nucleotide-long RNA fragments homologous to INGI and SLACS have been found to be very abundant (Djikeng *et al*. 2001). Thirtyone per cent of all siRNAs sequenced from these cells came from INGI and SLACS. Since the relative abundance of INGI and SLACS RNA in total cellular RNA is very low (INGI is 15-fold less and SLACS is 150-fold less than tubulin RNA), the relative level of siRNAs for them is very high. It is speculated that siRNAs corresponding to INGI and SLACS serve as RNAi to downregulate the activity of these elements. If they are involved in RNAi, they should be derived from dsRNA. Therefore INGI and SLACS need to be transcribed from both strands. In case of INGI it is known that transcripts originate from both strands (Mur phy *et al*. 1987); for SLACS there is no information. Evidence from *Caenorhabditis elegans* shows that RNAi can cause germ-line silencing of transposons (Ketting *et al*. 1999). RNAi may emerge as a general pathway to regulate the level of transposition of transposable elements.

Transposons—parasitic DNA or useful allies?

Transposons are considered to be ‘selfish’ genetic elements which have a transmission advantage relative to other parts of the genome, but are either neutral or detrimental to the organism’s fitness (Hurst and Werren 2001). For this reason it is thought that transposons would go to fixation in sexually reproducing species but are unlikely to persist in asexual organisms. This view is supported by the finding that the Bdelloid rotifer, an ancient asexual taxon, completely lacks retrotransposons, although it was positive for DNA transposon sequences (Arkhipova and Meselson 2001). The discovery of three retrotransposon families in *G. lamblia*, a species thought to be asexual, is in apparent contradiction of the above. However, two of the retrotransposons in *G. lamblia* are probably beneficial to their host by being subtelomeric in location; the third family seems to be totally inactivated by mutations. *E. histolytica* also reproduces asexually, and there is no evidence of a sexual cycle. Its genome is invaded by at least one (and probably more) pair of LINE-like and SINE-like autonomous and nonautonomous retrotransposons. Although most copies of these elements are mutated, the possibility of a few active copies exists. Further work will show whether the active copies are sequestered to ‘safe’ locations in the genome, or whether *E. histolytica* is indeed capable of sexual reproduction.

The location of transposons at subtelomeric regions (e.g. in *G. lamblia*) or at the telomeres themselves (e.g. in *D. melanogaster*) shows that transposons are sometimes coopted by the genome to serve a beneficial purpose. Phylogenetic analysis indicates that telomeres and transposons may have an ancient link. Telomerase, the enzyme that synthesizes telomeres, has an RT component that is structurally similar to the RT of non-LTR retrotransposons. Therefore, retroelements may have evolved from telomerase genes or vice versa (Hurst and Werren 2001).

A number of retrotransposons observed in protozoan parasites insert site-specifically in the SL-RNA genes. While non-site-specific elements may have a role in introducing variations that effectively reshuffle the genome and may influence the rate of evolutionary development, such a role cannot be ascribed to site-specific insertion elements. Is there a functional role for these elements, or have they evolved the ability of site-specific insertion into a multicopy gene to ensure their own survival? It is possible that these elements may help in the maintenance of SL-RNA genes by amplifying and disseminating them. So far there has been no experimental demonstration of actively transposing copies of retrotransposons in parasites. One hopes that future experiments will reveal the evolutionary purpose of mobile genetic elements in the genomes of protozoan parasites.

Acknowledgements

This work was supported by a grant from the Council of Scientific and Industrial Research and the Department of Science and Technology. A. B. is grateful for a Junior Research Fellowship from the Council of Scientific and Industrial Research.

References

Adam R. D. 1991 The biology of *Giardia* sp. *Microbiol. Rev.* **55**, 706–732.

Agabian N. 1990 Trans splicing of nuclear pre-mRNAs. *Cell* **61**, 1157–1160.

Aksoy S., Lalor T. M., Martin J., Van der Ploeg L. H. T. and Richards F. F. 1987 Multiple copies of a retroposon interrupt spliced leader RNA gene in the African trypanosome *Trypanosoma gambiense*. *EMBO J*. **6**, 3819–3826.

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

Arkhipova I. and Meselson M. 2001 Transposable elements in sexual and ancient asexual taxa. *Proc. Natl. Acad. Sci. USA* **97**, 14473–14477.

Arkhipova I. R. and Morrison H. G. 2001 Three retrotransposon families in the genome of *Giardia lamblia*: Two telomeric, one dead. *Proc. Natl. Acad. Sci. USA* **98**, 14497–14502.

Bagchi A. 2001 Studies on structure and organization of chromosomes in *Entamoeba histolytica*. Ph.D. thesis, Jawaharlal Nehru University, New Delhi, India.

Bagchi A., Bhattacharya A. and Bhattacharya S. 1999 Lack of a chromosomal copy of the circular rDNA plasmid of *Entamoeba histolytica*. *Int. J. Parasitol*. **29**, 1775–1783.

Bellofatto V., Cooper R. and Cross G. A. M. 1988 Discontinuous transcription in *Leptomonas seymouri*: presence of intact and interrupted mini-exon gene families. *Nucl. Acids Res*. **16**, 7437–7456.

Boeke J. D. 1997 Lines and Alus—the polyA connection. *Nat. Genet*. **16**, 6–7.

Boeke J. D. and Corces V. G. 1989 Transcription and reverse transcription of retrotransposons. *Annu. Rev. Microbiol*. **43**, 403–434.

Bringaud F., Biteau N., Melville S. E., Hez S., El-Sayed N. M., Leech V. *et al*. 2002 A new, expressed multigene family containing a hotspot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. *Euk. Cell* **1**, 137–151.

Burke W. D., Eickbush D. G., Xiong Y., Jakubczak J. and Eickbush T. H. 1993 Sequence relationships of retrotransposable elements R1 and R2 within and between divergent insect species. *Mol. Biol. Evol*. **10**, 163–185.

Cano M. I., Gruber A., Vazquez M., Cortes A., Levin M. J., Gonzalez A. *et al*. 1995 Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. *Mol. Biochem. Parasitol*. **71**, 273–278.

Chiurillo M. A., Cano I., Da Silveira J. F. and Ramirez J. L. 1999 Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite *Trypanosoma cruzi*. *Mol. Biochem. Parasitol*. **100**, 173–183.

Cost G. J. and Boeke J. D. 1998 Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37**, 18081–18093.

Cruz-Reyes J., Ur-Rehman T., Spice W. M. and Ackers J. P. 1995 A novel transcribed repeat element from *Entamoeba histolytica*. *Gene* **166**, 183–184.

Diamond L. S. and Clark C. G. 1993 A redescription of *Entamoeba histolytica* Schaudinn 1903 (emended Walker 1911) separating it from *Entamoeba dispar* Brumpt 1925. *J. Euk. Microbiol*. **40**, 340–344.

Djikeng A., Shi H., Tschudi C. and Ullu E. 2001 RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA* **7**, 1522–1530.

El-Sayed N. M., Hegde P., Quackenbush J., Melville S. E. and Donelson J. E. 2000 The African trypanosome genome. *Int. J. Parasitol*. **30**, 329–345.

Finnegan D. J. 1989 Eukaryotic transposable elements and genome evolution, *Trends Genet*. **5**, 103–107.

Finnegan D. J. 1997 Transposable elements: How nonLTR retrotransposons do it. *Curr. Biol*. **7**, R245–R248.

Gabriel A. and Boeke J. D. 1991 Reverse transcriptase encoded by a retrotransposon from the trypanosomatid *Crithidia fasciculata*. *Proc. Natl. Acad. Sci. USA* **88**, 9794–9798.

Gabriel A., Sisodia S. S. and Cleveland D. W. 1987 Evidence of discontinuous transcription in the trypanosomatid *Crithidia fasciculata*. *J. Biol. Chem*. **262**, 16192–16199.

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

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

Hay J. M., Jones M. C., Blakebrough M. L., Dasgupta I., Davies J. W. and Hull R. 1991 An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. *Nucl. Acids Res*. **19**, 2615–2621.

Hurst G. D. D. and Werren J. H. 2001 The role of selfish genetic elements in eukaryotic evolution. *Nat. Rev. Genet*. **2**, 597–606.

Hutchison C. A., Hardies S. C., Loeb D. D., Shehee W. R. and Edgell M. H. 1989 LINEs and related retroposons: long interspersed repeated sequences in the eukaryotic genome. In *Mobile DNA*. (ed. D. E. Berg and M. M. Howe), pp. 593–617. American Society for Microbiology, Washington, D. C.

Johnson M. S., McClure M. A., Feng D. F., Gray J. and Doolittle R. F. 1986 Computer analysis of retroviral Pol genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. *Proc. Natl. Acad. Sci. USA* **83**, 7648–7652.

Kazazian H. H. 1998 Mobile elements and disease. *Curr. Opin. Genet. Dev*. **8**, 343–350.

Kazazian H. H. 2000 L1 retrotransposons shape the mammalian genome. *Science* **289**, 1152–1153.

Ketting R. F., Haverkamp T. H., van Luenen H. G. and Plasterk R. H. 1999 Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNase D. *Cell* **99**, 133–141.

Kimmel B. E., ole-Moiyoi O. K. and 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.

Levskaya O., Slot F., Pavlova M. and Pardue M. L. 1994 Structure of the *Drosophila* HET-A transposon—a retrotransposon-like element forming telomeres. *Chromosoma* **103**, 215–224.

Lodes M. J., Smiley B. L., Stadnyk A. W., Bennett L., Myler P. J. and Stuart K. 1993 Expression of a retroposon-like sequence upstream of the putative *Trypanosoma brucei* variant surface glycoprotein gene expression site promoter. *Mol. Cell. Biol*. **13**, 7036–7044.

Loeb D. D., Padgett R. W., Hardies S. C., Shehee W. R., Comer M. B., Edgell M H. and Hutchison C. A. 1986 The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons. *Mol. Cell. Biol*. **6**, 168–182.

Luan D. D., Korman M. H., Jakubczak J. L and Eickbush T. H. 1993 Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for nonLTR retrotransposition. *Cell* **72**, 595–605.

Lushbaugh W. B. and Miller J. H. 1988 The morphology of *Entamoeba histolytica*. In *Amebiasis: human infection by Entamoeba histolytica* (ed. J. I. Ravdin), pp. 41–68. Wiley, New York.

Malik H. S. and Eickbush T. H. 2000 NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* **154**, 193–203.

Malik H. S., Burke W. D. and Eickbush T. H. 1999 The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16**, 793–805.

Martin F., Maranon C., Olivares M., Alonso C. and Lopez M. C. 1995 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. *J. Mol. Biol.* **247**, 49–59.

Martin S. L. and Bushman F. D. 2001 Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **21**, 467–475.

Miller J., McLachlan A. D. and Klug A. 1985 Repetive zinc binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J.* **4**, 1609–1614.

Mizrokhi L. J., Georgieva S. G. and Ilyin Y. V. 1988 Jockey, a mobile *Drosophila* element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. *Cell* **54**, 685–691.

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

Okada N., Hamada M., Ogiwara I. and Ohshima K. 1997 SINEs and LINEs share common 3′ sequences: a review. *Gene* **205**, 229–243.

Olivares M., Alonso C. and Lopez M. C. 1997 The open reading frame 1 of the L1Tc retrotransposon of *Trypanosoma cruzi* codes for a protein with apurinic-apyrimidinic nuclease activity. *J. Biol. Chem.* **272**, 25224–25228.

Olivares M., Thomas C. M., Alonso C. and Lopez M. C. 1999 The L1Tc, long interspersed nucleotide element from *Trypanosoma cruzi*, encodes a protein with 3′-phosphatase and 3′-phosphodiesterase enzymatic activities. *J. Biol. Chem.* **274**, 23883–23886.

Pardue M. L., Danilevskaya O. N., Lowenhaupt K., Wong J. and Erby K. 1996 The gag coding region of the *Drosophila* telomeric retrotransposon, HeT-A, has an internal frame shift and a length polymorphic region. *J. Mol Evol.* **43**, 572–583.

Pardue M., Baryshe P. G. and Lowenhaupt K. 2001 Another protozoan contributes to understanding telomeres and transposable elements. *Proc. Natl. Acad. Sci. USA* **98**, 14195–14197.

Sharma R., Bagchi A., Bhattacharya A. and Bhattacharya S. 2001 Characterization of a retrotransposon-like element from *Entamoeba histolytica*. *Mol. Biochem. Parasitol.* **116**, 45–53.

Tchurikov N. A., Gerasimova T. I., Johnson T. K., Barbakar N. I., Kenzior A. L. and Georgiev G. P. 1989 Mobile elements and transposition events in the cut locus of *Drosophila melanogaster*. *Mol. Gen. Genet.* **219**, 241–248.

Teng S.-C., Wang S. X. and Gabriel A. 1995 A new non-LTR retrotransposon provides evidence for multiple distinct site-specific elements in *Crithidia fasciculata* mini exon arrays. *Nucl. Acids Res.* **23**, 2929–2936.

Teng S.-C., Kim B. and Gabriel A. 1996 Retrotransposon reverse transcriptase-mediated repair of chromosomal breaks in *Saccharomyces cerevisiae*. *Nature* **383**, 641–644.

Vassella E., Roditi I. and Braun R. 1996 Heterogeneous transcripts of RIMR/INGI retroposons in *Trypanosoma brucei* are unspliced. *Mol. Biol. Parasitol.* **82**, 131–135.

Vazquez M., Schijman A. G. and Levin M. J. 1994 A short interspersed repetitive element provides a new 3′ acceptor site for trans-splicing in certain ribosomal P2-protein genes of *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* **64**, 327–336.

Vazquez M., Lorenzi H., Schijman A. G., Ben-Dov C. and Levin M. J. 1999 Analysis of the distribution of SIRE in the nuclear genome of *Trypanosoma cruzi*. *Gene* **239**, 207–216.

Vazquez M., Ben-Dov C., Lorenzi H., Moore T., Schijman A. and Levin M. J. 2000 The short interspersed repetitive element of *Trypanosoma cruzi*, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons. *Proc. Natl. Acad. Sci. USA* **97**, 2128–2133.

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

Weiner A. M. 2000 Do all SINEs lead to LINEs*? Nat. Genet.* **24**, 332–333.

Willhoeft U. and Tannich E. 1999 The electrophoretic karyotype of *Entamoeba histolytica*. *Mol. Biochem. Parasitol.* **99**, 41–53.

Willhoeft U., Heidrun B. and Tannich E. 1999 Analysis of cDNA expressed sequence tags from *Entamoeba histolytica*: Identification of two highly abundant polyadenylated transcripts with no overt open reading frames. *Protist* **150**, 61–70.

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