The non-LTR retrotransposons of *Entamoeba histolytica*: genomic organization and biology

Devinder Kaur1 · Mridula Agrahari2 · Alok Bhattacharya3 · Sudha Bhattacharya3

Received: 21 August 2021 / Accepted: 26 November 2021 / Published online: 9 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Genome sequence analysis of *Entamoeba* species revealed various classes of transposable elements. While *E. histolytica* and *E. dispar* are rich in non-long terminal repeat (LTR) retrotransposons, *E. invadens* contains predominantly DNA transposons. Non-LTR retrotransposons of *E. histolytica* constitute three families of long interspersed nuclear elements (LINEs), and their short, nonautonomous partners, SINEs. They occupy ~11% of the genome. The EhLINE1/EhSINE1 family is the most abundant and best studied. EhLINE1 is 4.8 kb, with two ORFs that encode functions needed for retrotransposition. ORF1 codes for the nucleic acid-binding protein, and ORF2 has domains for reverse transcriptase (RT) and endonuclease (EN). Most copies of EhLINEs lack complete ORFs. ORF1p is expressed constitutively, but ORF2p is not detected. Retrotransposition could be demonstrated upon ectopic overexpression of ORF2p, showing that retrotransposition machinery is functional. The newly retrotransposed sequences showed a high degree of recombination. In transcriptomic analysis, RNA-Seq reads were mapped to individual EhLINE1 copies. Although full-length copies were transcribed, no full-length 4.8 kb transcripts were seen. Rather, sense transcripts mapped to ORF1, RT and EN domains. Intriguingly, there was strong antisense transcription almost exclusively from the RT domain. These unique features of EhLINE1 could serve to attenuate retrotransposition in *E. histolytica*.

Keywords *Entamoeba histolytica* · Non-LTR retrotransposons · EhLINE1 · LINE transcription · LINE ORFs · Reverse transcriptase · Endonuclease · Antisense RNA

Introduction
Transposable elements (TEs) are capable of moving to new genomic locations, and are, thus, potent drivers of genomic evolution. They mediate various important events, such as gene inactivation, regulation of gene expression, genome expansion and sequence reorganization (Craig 2002; Kazazian 2000; Richardson et al. 2015). TEs are generally divided into two major groups depending on their mechanism of transposition (Finnegan 1989; Kazazian 1998). The DNA transposons transpose via a DNA intermediate. The retrotransposons move by reverse transcription of an RNA intermediate. They belong to two major classes: the long terminal repeat (LTR)-containing and the non-LTR-containing retrotransposons (Boeke and Corces 1989). The LTR retrotransposons share structural and functional similarity with retroviruses. Compared with the LTR retrotransposons, the non-LTR retrotransposons use a fundamentally different mechanism for transposition, called target-primed reverse transcription (Elbarbary et al. 2016; Finnegan 1997; Luan...
et al. 1993). Active copies of TEs that encode trans-acting functions required for transposition are relatively rare in most organisms, since these could potentially destabilize the genome by insertional mutagenesis and other mechanisms (Deininger and Batzer 1999). Most copies do not encode functional gene products due to multiple mutations throughout their sequences. Active copies of autonomous TEs that are capable of expressing the functions needed for transposition, also support the transposition of nonautonomous TEs. The latter contain requisite sequences recognized by the transposition machinery of their partner autonomous TE. Nonautonomous SINEs that transpose through the functions encoded by their partner LINEs share a stretch of sequence similarity at their 3′-ends that is possibly recognized by the LINE machinery (Boeke 1997; Kajikawa and Okada 2002; Okada and Hamada 1997).

A variety of TEs exist in parasite genomes, where they could potentially contribute to genomic changes associated with virulence phenotypes (Bhattacharya et al. 2002; Bringaud et al. 2008; Thomas et al. 2010). The trypanosomatids—Trypanosoma brucei, Trypanosoma cruzi and Crithidia fasciculata (Macías et al. 2018) contain a large variety of TEs, some of which insert site-specifically in the spliced-leader RNA genes, and others are dispersed (Bhattacharya et al. 2002). INGI/RIME are non-LTR retrotransposons of T. brucei that insert close to repetitive genes, including the variant surface glycoprotein genes (Bringaud et al. 2002; Kimmel et al. 1987). T. brucei contains another non-LTR retrotransposon SLACS inserted in the spliced leader genes (Aksoy et al. 1990). T. cruzi contains the non-LTR retrotransposon L1Tc (Martin et al. 1995), which is similar to INGI/RIME of T. brucei while the counterpart of SLACS in T. cruzi is CZAR (Villanueva et al. 1991). T. cruzi also contains an LTR retrotransposon SIRE/VIPER (Vázquez et al. 2000). Leishmania major contains non-LTR retrotransposons, LmSIDERs, that insert within the 3′-UTRs of mRNAs (Bringaud et al. 2007). Giardia lamblia contains three families of non-LTR retrotransposons. Two of these are located in subtelomeric regions (Arkipov and Morrison 2001).

Entamoeba histolytica, an early-branching protist, causes amebiasis, and is the second leading cause of morbidity and mortality due to parasitic disease in humans (Ralston and Petri 2011). The most abundant TEs in this organism are the non-LTR retrotransposons EhLINEs and EhSINEs which together occupy about 11% of the genome. Here we describe the structure and organization of these non-LTR retrotransposons and highlight some of their unique features, which include lack of full-length EhLINE transcripts, and massive antisense transcription of the reverse transcriptase (RT) coding region. These features are important as they could potentially limit active retrotransposition both by restricting the number of full-length EhLINE transcripts and by attenuating RT expression through antisense RNA.

Transposable elements in Entamoeba

Historical perspective

The first indication that retrotransposons exist in E. histolytica came from the analysis of a multicopy, 2.3 kb DNA (HMc) (Mittal et al. 1994). Its nucleotide sequence showed a very strong match with known reverse transcriptase sequences, especially those from non-LTR retrotransposons. Sequence analysis of overlapping, HMc-hybridizing clones from E. histolytica (Cruz-Reyes et al. 1995; Willhoef et al. 1999). Interestingly, it was found that 74 nucleotides at the 3′-end of IE/Ehapt2 were almost identical in sequence with the 3′-end of EhRLE1 (Bagchi 2001). Since the 3′-ends of many SINEs are homologous to the 3′- ends of their respective LINEs (Boeke 1997; Okada et al. 1997), it was speculated that EhRLE1 and IE/Ehapt2 could constitute a LINE/SINE pair. The subsequent initiation of the E. histolytica genome project by an international consortium (Loftus et al. 2005) led to the more precise definition of transposable elements in E. histolytica and other Entamoeba species.

Predominant classes of TEs in Entamoeba species

Whole-genome sequencing (WGS) of several Entamoeba species revealed the presence of TEs in their genomes. The first WGS data to be published was for E. histolytica strain HM-1:IMSS (Loftus et al. 2005). This was re-assembled and re-annotated, along with the assembly and annotation of the sequenced genomes of other important Entamoeba species; Entamoeba dispar, Entamoeba moshkovskii and Entamoeba invadens (Lorenzi et al. 2008; Pritham et al. 2005). E. histolytica is pathogenic in humans and causes amoebic dysentery and liver abscess. E. dispar is morphologically identical to E. histolytica, resides in the human gut, but is nonpathogenic. E. moshkovskii is free living. E. invadens is a parasite of reptiles. Unlike E. histolytica, E. invadens can form cysts in axenic cultures and is therefore used as a model system for encystation studies. Key features of the genomes of these species, and status of TEs are given in Table 1. Genome sequence analysis showed that E. histolytica and E. dispers have very few DNA transposons (Pritham et al. 2005), but have an abundance of retrotransposons that occupy 11% and 7% of their genomes,
respectively. Conversely, the genomes of \( E. \) *invadens* and \( E. \) *moshkovskii* contain few retrotransposons (0.1% of the genome) but have an abundance of DNA transposons. Amongst the retrotransposons, LTR retrotransposons could not be identified in any of the *Entamoeba* species examined. Neither the LTR-specific integrase nor RT sequence showed any matches. These results suggest the absence of LTR retrotransposons in these genomes (Pritham et al. 2005). All retrotransposons belonged to the non-LTR class.

DNA transposons of *Entamoeba*

Among *Entamoeba* species, *E. invadens* and *E. moshkovskii* are enriched with DNA transposons compared to retrotransposons. Based on transposase (Tpase) sequences, DNA transposons from these *Entamoeba* species belonged to the four superfamilies hAT, Mutator, Tcl/mariner, and piggyBac. DNA transposons constituted ~9.4% and 7% of these genomes respectively, with copy numbers of 673 in *E. invadens* and 723 in *E. moshkovskii* (Pritham et al. 2005; Lorenzi et al. 2008). Apart from high copy numbers, DNA transposons of *E. invadens* and *E. moshkovskii* also showed great sequence diversity. A total of 48 distinct DNA Tpase families were identified in the genome contigs of *E. invadens* (Mutator-8, Tcl/mariner-31, hAT-6, and piggyBac-3) (Pritham et al. 2005) compared to only 18 families in *D. melanogaster*. Analysis of TSDs flanking the DNA transposons showed that their length and/or sequence were characteristic of the respective superfamilies reported in eukaryotes (TA for Tcl/mariner, TTAA for piggyBac, 8 bp for hAT, and 9 bp for Mutator).

Of the four transposon superfamilies, sequences corresponding to two of them (Mutator and Tcl/mariner) were also found in *E. histolytica* and *E. dispar*. Using the most conserved region of each clade of Mutator Tpase (domain pfam00872) in TBLastN searches, Mutator-like Tpase sequences were identified in the four *Entamoeba* genomes. The copy numbers varied significantly for each species: 5 in *E. dispar*, 2 in *E. histolytica*, 198 in *E. invadens*, and 322 in *E. moshkovskii* (Pritham et al. 2005). All copies of the Mutator superfamily in *E. histolytica* and *E. dispar* belonged to the EMULE group, while in *E. invadens* and *E. moshkovskii* there was a second group, Phantom as well. Within the Tcl/mariner superfamily, five different families (pogo-like Piglet-Ei1, Fot1-like Gemini-Ei1, Tcl-like Hydragos-Em1, Mogwai Mogwai-Ei1, Gizmo-Ei1) were found in the genomes of *E. invadens* and *E. moshkovskii* (Pritham et al. 2005). Using an analysis tool, Transposon-PSI, sequences homologous to mariner were identified in *E. histolytica* and *E. dispar* as well. These were closely related to the Hydragos family (Lorenzi et al. 2008), while sequences for the other four families could not be identified.
Sequence analysis of DNA transposons in *Entamoeba* showed signs of recent transposition activity in several cases. The Mutator TEs of *E. invadens* had two copies inserted at different chromosomal positions, and flanked by different TSDs. These copies were 99.9% identical. They had apparently intact ORFs encoding potentially active enzymes, and had perfect terminal inverted repeats indicative of recent transposition. Gemini elements appear to be recently active in *E. moshkovskii* as evidenced by the presence of multiple nearly identical Tpase fragments. Many Hydargos Tpase fragments from *E. moshkovskii* are also nearly identical, suggesting recent amplification of these lineages in the genome.

The presence of transposons from Mariner and Mutator superfamilies in the four *Entamoeba* genomes strongly suggested that these sequences were present in the common ancestor and not acquired horizontally after the species diverged from each other.

**The *Entamoeba* repetitive elements (ERE)**

Apart from the canonical class I and class II TEs, another repetitive sequence ERE1 is found in the genomes of *E. histolytica*, *E. dispar*, and *E. invadens* where it occupies 4.9%, 2.3% and 0.4% of the genomes, respectively (Lorenzi et al. 2008). The ERE1 sequences were flanked by terminal inverted repeats (TIRs: SINE1/SINE3). In *E. dispar* the genomic organization of ERE1 and its flanking region showed that the central region of Ed_ERE1 and Ed_SINE1, were inserted alone throughout the *E. dispar* genome, suggesting that at some point during evolution, these two elements were able to transpose independently from each other (Lorenzi et al. 2008). In *E. invadens*, the EI_ERE1 lacked a large flanking TIR. The Eh_ERE1 of *E. histolytica* is 7,160 bp in length, and consists of a core region flanked by inverted repeats. The core region contains an open reading frame (ORF) that could encode a 369 amino acids protein. The ratio of non-synonymous to synonymous substitutions (Dn/Ds) for this gene in each *Entamoeba* species was computed which indicated that the gene coded by ERE1 is under purifying selection. Comparative sequence analysis revealed several syntenic regions between *E. histolytica* and *E. dispar* where ERE1 was inserted in one genome but not the other, supporting the hypothesis that at some point in the evolution of these species ERE1 might have been able to transpose. However, no homologue to the protein encoded by ERE1 was found in the database, and it did not show homology with proteins encoded by known TEs. Hence it is not clear whether this repeat sequence is, indeed, a TE. A second class of repeat sequence, ERE2 was found exclusively in *E. histolytica* (Lorenzi et al. 2008), and occupied 3.5% of the genome. It contained a putative ORF coding for a 173 aa polypeptide with no homology to any known protein. Unlike Eh_ERE1, not a single copy of ERE2 was found in the genomes of either *E. dispar* or *E. invadens*, suggesting that *E. histolytica* could have acquired this sequence independently after diverging from *E. dispar*, its closest relative.

**Phylogenetic analysis of *Entamoeba* transposable elements**

The non-LTR retroposons in the sibling species *E. histolytica* and *E. dispar* belong to three families each of LINEs and SINEs, with significant sequence relatedness (Bakre et al. 2005; Lorenzi et al. 2008; Pritham et al. 2005; Sharma et al. 2001; Willhoeft et al. 1999; Shire and Ackers 2007; Dellen et al. 2002; Huntley et al. 2010). Their key features, including genomic abundance, size and number of full-length/truncated copies are given in Table 2.

A consensus sequence of each EhLINE family was reconstructed manually by selecting the most common nucleotide at each position. All three EhLINE families had well-conserved RT and endonuclease (EN) domains in ORF2. However, ORF1 could not be identified in EhLINE3, possibly due to the accumulation of too many mutations (Bakre et al. 2005) (Fig. 1a). Phylogenetic analysis based on the EhLINE1 RT consensus sequence showed closest match with RTs of the R4 clade, most notably the R4 retrotransposon of *Ascaris lumbricoides* and Dong retrotransposon of *Bombyx mori* (Fig. 1a, b). Similarly, the LINE EN domain had sequence features resembling Type IIS restriction endonucleases, and was very similar to the domains in R2, R4, and CRE clades of non-LTR retrotransposons. In addition to sequence matches, the structural arrangement of protein-coding domains and motifs (zinc finger motifs, c-myb-like motifs) in *Entamoeba* LINEs also placed them in the R4 clade of non-LTR retroposons (Bhattacharya et al. 2002; Eickbush and Malik 2014; Dellen et al. 2002; Yang et al. 1999).

The RT consensus sequence was also used to determine the evolutionary interrelatedness of the Eh/Ed_LINE families, which showed that they were all derived from a single ancestral sequence. At a later point in evolution, LINE TEs split into two separate lineages giving rise to the Eh/EdLINE1 and Eh/EdLINE2 subfamilies. Subsequently, a third LINE subfamily diverged from LINE1 giving rise to EhLINE3 in *E. histolytica* and EdLINE3 in *E. dispar* (Lorenzi et al. 2008). The non-autonomous SINEs also already existed in the common ancestor before the speciation process that gave rise to *E. histolytica* and *E. dispar*. The most abundant EhSINE, EhSINE1, is closely related to EdSINE3, while the most abundant EdSINE, EdSINE1 is closely related to EhSINE3. The 5′ end of EdSINE1/EhSINE3 is more similar to EhSINE2 whereas its 3′ end resembles EhSINE1 (Fig. 1c). This suggests that in the common ancestor of *E. histolytica* and *E. dispar*, EdSINE1/EhSINE3 originated as a chimeric
SINE; derived from the precursor sequence of EhSINE2/EdSINE2 and EhSINE1/EdSINE3 (Lorenzi et al. 2008).

Ei_LINE is a 5043 bp sequence flanked by TSDs (Pritham et al. 2005). Just two complete copies of Ei_LINE were found in E. invadens and neither of them had a complete ORF coding for a reverse transcriptase protein. Phylogenetic analysis based on manual reconstruction of the reverse transcriptase consensus sequence indicated that all LINEs found in the three Entamoeba species (E. histolytica and E. dispar and E. invadens) were derived from a single ancestral sequence that was already present before they diverged from each other (Fig. 1b) (Lorenzi et al. 2008).

Phylogenetic analysis based on both RT and EN domains of non-LTR retrotransposons put E. histolytica and E. dispar in a single clade. Similarly, the phylogenetic profile of Entamoeba species based on DNA transposons using Entamoeba EMULE putative transposase domains showed that they form a distinct clade for EMULE sequences which includes at least five lineages (Pritham et al. 2005). Two lineages are specific to E. moshkovskii, and two lineages are specific to E. invadens. The remaining lineage includes sequences from E. dispar, E. histolytica, and E. moshkovskii. The phylogenetic relationships within this lineage are in good agreement with the species phylogenies established from ribosomal sequences (Silberman et al. 1999), suggesting that this lineage of EMULE was present in the common ancestor of these Entamoeba species and was vertically inherited (Pritham et al. 2005).

### Genomic distribution of EhLINEs/EhSINEs

Early Southern hybridization studies of electrophoretically separated chromosomes of E. histolytica using EhLINE1/EhSINE1 probes (Bagchi et al. 1999) showed that these retrotransposon sequences were widespread on all chromosomal bands, and did not seem to be located at chromosome ends. Computational analysis of 2 kb upstream and downstream regions flanking the insertion sites did not show conserved sequences. However, all the retrotransposon copies seemed to insert in AT-rich sequences, with a clear preponderance of T-residues in a 50-nt stretch upstream of the site of insertion of each element (Fig. 2). About 80% of the time there was either a gene or another retrotransposon copy located in the close vicinity of a given retrotransposon sequence. Clusters of retrotransposon copies were frequently found within 0.1 kb of each other. They were found in both orientations, and no clear preference of any pairs of retrotransposon sequences occurring near each other could be discerned (Bakre et al. 2005). TE repeat-clusters were also found at syntenic break points between E. histolytica and E. dispar and could serve as recombination hot spots (Lorenzi et al. 2008). Housekeeping genes were rarely found in the vicinity of EhLINEs/SINEs, and there was no instance of a retrotransposon sequence inserted within a gene (Bakre et al. 2005) (Fig. 2).

The contribution of EhLINE-encoded endonuclease (EN) to target site selection was investigated. The sequence-specificity of EN was tested in vitro, and a
consensus sequence, 5′-GCATT-3′, was shown to be efficiently nicked between A-T and T-T. The requirement for upstream G residue possibly served to limit retrotransposition in the AT-rich *E. histolytica* genome. The nicking specificity, although important, probably plays only a limited role in insertion site selection. The importance of unique DNA structure at insertion sites has been shown for the bacterial TE Tn7 (Kuduvalli et al. 2001) and the *D. melanogaster* P element (Liao 2000). Computational analysis of EhSINE1 pre-insertion loci in *E. histolytica* showed unique features based on DNA structure, thermodynamic considerations and protein/nucleosome interaction measures. Target sites could readily be distinguished from other genomic sites based on these criteria. The -10 to -35 region of a true insertion site tended to be rigid as indicated by propeller twist and bendability measures, and could melt easily (Mandal et al. 2006). Thus, a combination of favorable DNA structure and preferred EN nicking sequence may determine the genomic hotspots for retrotransposition of EhLINEs/SINEs.

**EhLINE1**: the most abundant non-LTR retrotransposon of *E. histolytica*

The only *Entamoeba* TE whose retrotransposition properties are understood in some detail is the *E. histolytica* LINE, EhLINE1. Amongst the three families of EhLINES, EhLINE1, present in an estimated 742 copies is the most abundant (Table 2). Among EhSINES, EhSINE1, the non-autonomous partner of EhLINE1, present in 493 copies is the most abundant. There are 742 copies of EhLINE1 in the size range of 42–4811 bp, with most copies in the range of 1–2 kb. There are only 61 full-length copies; the rest being truncated at one or both ends, or containing large internal deletions that are flanked by direct repeats (Kaur et al. 2021).

A typical full-length copy of EhLINE1 (4.8 kb) is composed of two non-overlapping ORFs. Analysis of the domain structures of EhLINE1-encoded ORFs shows that they share some similarities with LINE-encoded ORFs from other organisms. ORF1 (1.5 kb) contains an RNA recognition motif with single-stranded nucleic acid binding activity.
Fig. 2 Features of EhLINE/SINE genomic insertion sites. The insertion sites lacked a clear consensus sequence; however, conserved features were seen at most sites, as listed. Except for DNA structural features, other listed patterns were observed downstream of insertion sites as well. No instance of a copy inserting within a gene was observed. Most copies had the conserved sequence AGAT at the 5′-end.

| EhLINE/SINE upstream Region | Features                                                                                                                                 |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| -50 bp                      | T-rich region. (Bakre et al. 2005)                                                                                                                                                                   |
| -50 bp                      | Unique DNA structural features: Low bendability, Propeller twist, Stacking energy, Free Energy Profile, DNA denaturation energy, Protein induced deformability, Nucleosomal related features: bending energy and nucleosomal positioning profile (Mandal et al. 2006) |
| -100 bp                     | Presence of other EhLINE/SINE copies: indicative of clusters (Bakre et al., 2005; Lorenzi et al. 2008)                                                                                              |
| -500 bp                     | Frequent presence of genes: Mostly hypothetical; annotated genes include kinases, GTPases, heat shock proteins, and BspA. (Bakre et al., 2005; Lorenzi et al. 2008)                               |
| 5′-GCATT-3′                 | Consensus sequence for EhLINE1_EN-specific nick for element insertion. (Mandal et al. 2006)                                                                                                          |

Fig. 3 The full-length EhLINE1. A consensus sequence of the 4804 bp full-length EhLINE1 was reconstructed manually by selecting the most common nucleotide at each position (Bakre et al. 2005). The functional domains that could be identified by sequence analysis and biochemical assays are shown for ORF1 (adapted from Gaurav et al. 2017) and ORF2 (adapted from Mandal et al. 2004).
whereas ORF2 (~3 kb) codes for protein with reverse transcriptase (RT) and endonuclease (EN) domains (Fig. 3) (Gaurav et al. 2017; Mandal et al. 2004; Yadav et al. 2009). Most copies of EhLINE1 have deletions or other mutations, with intact ORFs being found in very few copies (Kaur et al. 2021). *E. histolytica* cells express ORF1p constitutively. However, ORF2 polypeptide could not be detected.

**The activities encoded by EhLINE1**

The accepted mechanism by which a LINE is thought to retrotranspose to a new genomic location is called target-site primed reverse transcription (TPRT), first proposed for the *Bombyx mori* LINE, R2Bm (Luan et al. 1993) and shown to also operate in human LINE-1 (Hancks and Kazazian 2012; Richardson et al. 2015). Broadly, TPRT is initiated by transcription of a full-length LINE copy, which is translated into a single polypeptide (in R2Bm) or two polypeptides, ORF1p and ORF2p (in human LINE-1). The polypeptides bind to their mRNA forming a ribonucleoprotein (RNP) particle. In human LINE-1 the RNP consists of multiple ORF1p trimers and only one or a few copies of ORF2p (Alisch et al. 2006; Doucet et al. 2010). The RNP is transported to the nucleus where the EN domain of ORF2p makes a nick at the new genomic target site. This provides a 3’ hydroxyl group that is used as a primer for reverse transcription of LINE RNA by the ORF2p RT domain (Cost et al. 2002). This is followed by cleavage of the target site on the opposite strand, and second-strand synthesis, following which a LINE copy is integrated into a new site (Christensen and Eickbush 2005). Thus, active retrotransposition requires the expression of functional LINE-encoded polypeptides, ORF1p and ORF2p.

**EhLINE1 ORF1p**

The EhLINE1 ORF1 encodes a predicted polypeptide of 498 amino acids of 60.5 kDa. Unlike the EN and RT domains of ORF2p which are well conserved, the ORF1ps from LINEs of various organisms lack well-conserved domains. Nevertheless, they do possess conserved functions like nucleic acid binding, nucleic acid chaperone and self-interaction properties (Kolosha and Martin 1997; Martin and Bushman 2001; Nakamura et al. 2012). Many of them also contain domains such as coiled coil needed for the formation of multimers, RNA recognition motif, and domains with basic regions, which help in the formation of ribonucleoprotein particles (Khazina and Weichenrieder 2009). Amongst the retrotransposons in the R2 group to which EhLINEs belong, the R2Bm N-terminal domain contains a DNA-binding region with Zinc finger motifs, and Myb domain, along with some RNA-binding residues upstream of RT domain (Christensen et al. 2005; Jamburuthugoda and Eickbush 2014). In EhLINE1 ORF1p, well-conserved functional domains could not be readily found by sequence search. RNA-binding residue prediction tools could detect a long RNA-binding stretch at the N-terminus (Gaurav et al. 2017). A coiled-coil domain was readily detected near the C-terminus. EhLINE1 ORF1p could not bind DNA, but could efficiently bind RNA as measured by electrophoretic mobility shift assay. The Kd values for RNA binding in the in vitro assay (Gaurav et al. 2017) were similar to reported values from other systems (Kolosha and Martin 2003). As predicted by the domain analysis, N-terminus of EhLINE1 ORF1p had nucleic acid binding activity, while the C-terminus appeared to promote the formation of multimeric complexes.

In an in vitro assay, the protein showed a comparable binding affinity with a EhLINE/SINE-specific RNA (derived from their conserved 3’-end), or an unrelated RNA (rRNA) (Gaurav et al. 2017). However, the ORF1p is expected to form RNP specifically with LINE/SINE RNAs in vivo. Possibly, specificity could be imparted by the tertiary structure of RNA in vivo or by interaction with ORF2p, or another cellular protein. ORF1p and LINE RNAs could also interact co-translationally, resulting in the cis-preference observed for ORF1p for its own RNA (Callahan et al. 2012).

**The EhLINE1 ORF2p-encoded endonuclease**

The endonuclease (EN) encoded by EhLINEs resembles type II-S restriction enzymes, and contains the conserved motifs CCHC, PDX12–14D, RHD and KXXX (Mandal et al. 2004), as seen in retrotransposons of the R4 clade (Eickbush and Malik 2014; Yang et al. 1999) (Fig. 3). This enzyme also has similarity with archaeal Holliday junction resolvases, which are implicated in second-strand DNA synthesis during LINE integration (Khadgi et al. 2019). The nicking activity of EhLINE1 endonuclease was demonstrated with supercoiled pBluescript plasmid DNA, using the purified recombinant protein. Further, the enzyme activity was tested on a native substrate. For this, an empty target site was located by analyzing homologous regions of the tetraploid *E. histolytica* genome, some of which may harbor the retrotransposon while others could be unoccupied. This exercise was done for EhSINE1 due to its small size, and one unoccupied site was identified (Mandal et al. 2004). A 176-bp fragment that contained this empty site was nicked by EN on the bottom strand at the precise point where EhSINE1 had inserted, showing that the enzyme retained its nicking specificity in vitro. The ability of EhLINE1 EN to nick an empty target site used for EhSINE1 insertion suggested that the nonautonomous EhSINE1 could, indeed, use the EhLINE1 EN for its own insertion.

EhLINE1 EN shared similarities with bacterial restriction endonucleases; its PDX12–14D metal-binding motif was required for activity, and upon the addition of DNA there...
was a conformational change (Yadav et al. 2009). However, unlike them, it lacked strict sequence specificity of the target site (Mandal et al. 2006; Yadav et al. 2009). It is possible that the EN domain was acquired from bacteria through horizontal gene transfer. Subsequently, nicking specificity may have been relaxed to select for retrotransposons that could spread to intergenic regions of the E. histolytica genome (Yadav et al. 2009). That horizontal gene transfer might have contributed significantly to E. histolytica genome evolution was suggested from analysis of the genome sequence which showed signatures of horizontal gene transfer. This was especially so in genes involved in carbohydrate and amino acid metabolism, contributing to the unique metabolic capabilities of this parasite (Loftus et al. 2005).

Phylogenetic analysis has shown that the R4 clade which includes EhLINEs, is amongst the oldest lineages of non-LTR retrotransposons and is closely related to some group II introns of mitochondria and bacteria (Yang et al. 1999). The site-specificity of EN domain is well-suited for group II introns which move in the very compact genomes of bacteria and mitochondria. As the same domain was acquired by non-LTR retrotransposons in eukaryotic genomes, it retained site specificity in some species like Bombyx mori and nematodes (Burke et al. 2002), but lost it in others like E. histolytica.

The ORF2p RT domain

EhLINE1 ORF2p has a well-conserved RT domain (Fig. 3) which shows best match with RTs of non-LTR retrotransposons of the R4 clade. In a sequence alignment of a large number of RT sequences from LTR- and non-LTR-retrotransposons, seven peptide regions (domains I–VII) were found to be highly conserved in all RTs (Xiong and Eickbush 1990). Of the 42 conserved residues in these domains, the EhLINE RT sequence differed in just three positions which move in the very compact genomes of bacteria and mitochondria. As the same domain was acquired by non-LTR retrotransposons in eukaryotic genomes, it retained site specificity in some species like Bombyx mori and nematodes (Burke et al. 2002), but lost it in others like E. histolytica.

Demonstration of de novo retrotransposition by ectopic expression of EhLINE1 ORF2

The majority of LINE copies in most organisms are inactive due to the accumulation of multiple mutations (Brouha et al. 2003). The human L1 system was first adapted for the study of de novo retrotransposition. A full length L1 retrotransposon copy lacking any mutations was expressed as an episome in cultured mammalian cells and retrotransposition could be demonstrated (Moran et al. 1996) not only of the L1 but also SINEs, such as Alu and SVA (Dewannieux et al. 2003; Hancks et al. 2011; Raiz et al. 2012). This system has been extensively used to gain insights into various aspects of mammalian LINE expression (Alish et al. 2006; Doucet et al. 2010; Feng et al. 1996; Kulpa et al. 2006; Martin et al. 2005; Soifer and Rossi 2006; Stenglein and Harris 2006; Yu et al. 2001). A cell-line assay to study the process of Entamoeba LINES/SINEs retrotransposition has been described, which has shown some unique features of the EhLINE1 system (Yadav et al. 2012).

Construction of E. histolytica cell-line to demonstrate retrotransposition

It has been found that cultured E. histolytica cells express ORF1p constitutively, but the ORF2 polypeptide is undetectable. In the absence of ORF2p, retrotransposition is unlikely to take place. To study retrotransposition a copy of EhLINE1 was reconstructed to remove stop codons from endogenous sequences, and E. histolytica cell line was constructed that could ectopically express ORF2p from a tetracycline (tet)-inducible promoter. The same cell line also expressed a EhSINE1 copy that contained a GC-rich tag, and a target site sequence, taken from the E. histolytica genome, where a SINE copy is known to insert. The latter sequence was contained in a 176-bp fragment (Mandal et al. 2004). This system was used to score retrotransposition occurring at the target site (Fig. 4a, d) (Yadav et al. 2012). Following tet addition, retrotransposition of the marked SINE copy to the target site was scored by a PCR-based assay using primers flanking the site. Specific amplicons indicative of SINE mobilization were obtained only in the presence of tet, and were dependent on ORF2p expression (Fig. 4b). The presence of target site duplications (TSD) was tested, which are indicative of true retrotransposition events. The mobilized copies were found to be flanked by a 22-bp TSD, whose sequence matched exactly with the TSD found at this insertion hot-spot at its genomic location (Fig. 4c) (Mandal et al. 2004). Since mobilization was seen only after ORF2p expression, and the insertion was seen at the retrotransposition hot-spot, with 22-bp TSDs, the scored events were very likely to be a result of retrotransposition rather than other processes like DNA recombination (Yadav et al. 2012). This demonstration of retrotransposition in an early branching eukaryote was notable, and indicated that the retrotransposition machinery in E. histolytica was otherwise intact, but kept in check by non-availability of ORF2p.

In vivo retrotransposition of human LINE1 (Moran et al. 1996) and SINEs (Dewannieux et al. 2003; Hancks and Kazazian 2012; Raiz et al. 2012) in cultured cells, and a LINE/SINE pair from eel (Kajikawa and Okada 2002), have been well documented. However, these studies did not analyse whether the retrotransposed copies had undergone any
sequence changes. The *E. histolytica* study tested whether any sequence changes were seen in the marked SINE copy as a result of retrotransposition (Yadav et al. 2012), which showed evidence of sequence changes resulting from recombination.

**Newly retrotransposed SINEs in *E. histolytica* show evidence of recombination**

Retrotransposed SINE copies at the target site were PCR amplified and the amplicons were cloned and sequenced. Of twenty-three randomly selected clones ten had sequences matching completely with the marked SINE. Eight of them lacked the tag and showed identity with genomic SINE copies rather than the tagged SINE. These were due to mobilization of the transcribed EhSINE1 copies in *E. histolytica* (Huntley et al. 2010). Interestingly, five sequences contained the 25-bp tag at the expected location but had low overall sequence identity with the marked SINE. Rather, the sequence on either side of the tag matched with 98–100% identity with different genomic SINE sequences (Yadav et al. 2012), indicative of recombination (Fig. 4e). Appropriate controls were used to show that the tag was unlikely to be acquired by the genomic SINE copies through a DNA-recombination event prior to retrotransposition. Rather, the recombinant SINEs were formed consequent to retrotransposition. Reverse transcriptase mediated recombination events leading to chimeric molecules have been reported in a number of systems, including in yeast Ty retrotransponson (Derr and Strathern 1993), in retroviruses where two co-packaged

---

**Fig. 4** Demonstration of retrotransposition in ORF2p over-expressing cell-line. a A retrotransposition-competent *E. histolytica* cell line was generated by transfecting with two plasmids. pEh-ORF2 provided ORF2p through a tetracycline (tet)-inducible promoter. pEh-SINE1 provided a EhSINE1 copy marked with a GC-rich tag (purple box), and a genomic sequence containing a known EhSINE1 insertion hotspot. b Retrotransposition events were scored by PCR amplification using primer A1 from the GC-rich tag and A2 from the insertion site in pEh-SINE1. Successful retrotransposition resulted in 435 bp amplicon, which was seen only in Eh-ORF2p expressing cells in the presence of tet. c The flanking sequence of retrotransposed copies showed the exact 22 bp target site duplication (TSD) present at the known genomic insertion site. d The newly retrotransposed EhSINE1 copies were PCR amplified, and 23 copies were sequenced. 10 sequences matched with the SINE copy in pEh-SINE1; 8 sequences matched with genomic EhSINE1 copies that had got mobilized; 5 sequences were recombinants of multiple SINE copies and had retained the tag. e A model to show the possible generation of recombinant SINEs by template switching of reverse transcriptase during retrotransposition (adapted from Yadav et al. 2012)
RNAs in the virion engage in recombination during reverse transcription (Delviks-Frankenberry et al. 2011), in non-LTR retrotransposons, where tripartite chimeras are found in a fungal genome (Gogvadze et al. 2007), and U6/L1 pseudogene chimeras in mammals (Garcia-Perez et al. 2007). Recombinants could result from the ability of RT enzyme to engage in template jumping by displacing the RNA template (Bibello and Eickbush 2002; Bibillo and Eickbush 2004), which could lead to the observed recombinants in E. histolytica SINEs as well. Such events could also be a source of sequence diversity in retropseudogenes.

**Transcriptomic analysis of all genomic copies of EhLINE1**

LINE copies in most organisms contain multiple mutations, including point mutations and large deletions. Expression of intact copies from constitutive promoters is further controlled by a variety of factors. These include DNA methylation, chromatin environment, transcription factor availability and post-transcriptional regulation (Crichton et al. 2014; Iwasaki et al. 2015; Reik 2007). Regulation of TE transcription is thus a highly complex process.

EhLINE1 transcription has been studied using different approaches. RNA-Seq analysis was done to identify the expressed and silent copies (Kaur et al. 2021). RNA-Seq data were analyzed using the expectation-maximization tool RSEM. Credibility intervals were used to resolve any uncertainties in assigning reads to individual copies. Of the 742 EhLINE1 copies, only 41 were transcriptionally active. Of the expressed copies, 21 had internal deletions/end truncations, and 20 were full-length.

Analysis of the transcription and transcriptomic data of EhLINE1 has revealed some interesting features, as described below.

**The promoter of EhLINE1**

LINEs in model systems are generally transcribed from a RNA polymerase II-driven internal promoter which is located at the LINE 5′-end. A single polycistronic transcript is produced that copies the entire LINE (Alexandrova et al. 2012; Minakami et al. 1992; Mizrokhi et al. 1988; Swergold 1990). The internal promoter (Pr77) of the T. cruzi LINE, L1Tc, is located at the 5′-end within 77 bp, and amongst protozoan parasites it is the best characterized (Heras et al. 2007; Macías et al. 2016). In EhLINE1 an internal promoter was identified within 200 bp at the 5′-end using a luciferase reporter assay. No second promoter activity was observed upstream of the RT domain, including the spacer region (Fig. 5) (Kaur et al. 2021).

**DNA methylation status of EhLINE1**

In mammals, a principle mechanism for retrotransponson silencing is transcriptional repression through DNA methylation on cytosine residues in the context of CpG dinucleotides (Hackett et al. 2012; Yoder et al. 1997). RNA-Seq analysis of EhLINE1 showed that only 41 out of 742 genomic copies were transcriptionally active. To check whether cytosine DNA methylation was involved in transcriptional silencing, methylation status was checked by bisulfite treatment of total genomic DNA, which converts unmethylated cytosines to uracil in DNA, while methylated cytosines are protected (Fraga and Esteller 2002). Two EhLINE1 copies—one silent and the other actively transcribed, were selected for sequence analysis following bisulfite treatment. The data showed that none of the cytosine residues (including the CpGs) were methylated in either of the two copies (Agrahari et al. 2017). As a control, the Hsp70 promoter was fully methylated in the same samples, as expected (Fisher et al. 2006). Further, the methylation status of a few selected cytosines in all 5′-intact EhLINE1 copies was checked by single nucleotide incorporation opposite cytosine in bisulfite-treated DNA, where dGTP would be incorporated if the cytosine was methylated. No evidence of cytosine methylation was found (Agrahari et al. 2017) indicating that the expression status of this retrotransposon was not correlated with promoter DNA methylation.

In another study, the DNA methylation status of RT sequences from *E. histolytica* and *E. invadens* was determined by sequence analysis of anti-5 methylcytosine affinity-purified DNA (Harony et al. 2006). No methylation was seen in *E. histolytica*, while the *E. invadens* RT sequences were methylated. Moreover, RT transcripts were seen in *E. histolytica* but not in *E. invadens*. Thus it seems that in addition to the extremely low copy number of LINEs in *E. invadens* (only two complete copies), DNA methylation could further contribute to the lack of retrotransposon activity in *E. invadens*.

**No full-length EhLINE1 transcripts were detectable in *E. histolytica***

The RNA-Seq reads were mapped to each expressed copy to get an estimate of the length of EhLINE1 transcripts. The reads did not uniformly map to the entire length of EhLINE1. Rather they mapped mainly to three regions: (1) 5′-end to 1517 bp; (2) 2440 to 3791 bp; (3) 3849 to 3′-end. The first region covered the entire ORF1, the second region covered the RT domain, while the third region covered the EN domain (Fig. 5). This showed that steady state EhLINE1 transcripts corresponded mainly to the functional protein domains. Northern hybridization with a panel of probes spanning EhLINE1 confirmed the RNA-Seq data (Kaur et al. 2021).
et al. 2021). Full-length EhLINE1 transcript (4.8 kb) was not seen in northern blots. The ORF1- and RT-probes hybridized with ~1.5 kb bands (Kaur et al. 2021).

Although 20/41 transcribed EhLINE1 copies were full-length, yet full-length transcripts were not detected. In human L1 retrotransposon many truncated transcripts correspond to internal polyadenylation sites (Perepelitsa-Belancio and Deininger 2003). EhLINE1 was searched for polyadenylation signals known to be used by E. histolytica genes (Hon et al. 2013). These were found at the end of ORF1 and ORF2. Polyadenylation signal was also found at the end of the RT domain although the T-rich stretch surrounding the polyadenylation site was poorly defined (Kaur et al. 2021). From the available data, it appears that EhLINE1 ORF1 could be transcribed from the promoter at 5′-end, and could be polyadenylated at the motif present at 3′-end of the ORF. The 1.5 kb RT transcript could be processed from a polycistronic transcript, or it could be transcribed from a second internal promoter (not detected so far), located upstream of the RT domain.

Read-through transcription or the use of upstream gene promoters did not seem to contribute to the expression of EhLINE1 copies. There was also no evidence of EhLINE1 transcripts originating from splicing involving transcripts of neighbouring genes (Kaur et al. 2021).

The RT region of EhLINE1 is transcribed massively in the antisense direction

Antisense transcription from LINEs is not uncommon, although the extent of sense transcription is generally greater (Criscione et al. 2016; Denli et al. 2015; Speek 2001). Strand-specific transcription of EhLINE1 was checked by RNA-Seq. Sense transcripts, as expected,
originated from 3 regions of EhLINE1 (ORF1, RT and EN) but antisense transcripts were coded almost exclusively from RT region (Fig. 5). Three full-length copies of EhLINE1 had intact ORF1 and were transcriptionally active. Sense strand expression was seen from these copies, but there were no antisense ORF1 transcripts. Only one full-length EhLINE1 copy had an intact ORF2 reading frame. RT antisense transcripts mapped maximally to this copy (Kaur et al. 2021). Strand-specific probes were used to validate antisense transcription by northern analysis. The ORF1 antisense probe gave no signal, whereas the antisense RT probe showed a strong signal of 1.5 kb (Fig. 4). Presence of antisense RT transcripts was further confirmed by checking the EST database. Two ESTs mapped in antisense orientation, and both matched with sequences in the RT domain. Sequence analysis of the antisense RNA showed that it was unlikely to encode any functional peptides/proteins. The presence of RT-specific antisense RNA could explain why E. histolytica cells expressed ORF1p at high levels, but ORF2p was undetectable (Yadav et al. 2012).

Whether antisense RNA has a role in attenuating the translation of RT domain remains to be studied. Such a role has been suggested for the LTR retrotransposon Ty1 in S. cerevisiae. Antisense RNAs of sizes between 0.5 and 1.0 kb, mapping to the Gag region of Ty1, act post-transcriptionally and inhibit reverse transcription by preventing the accumulation of mature Pol proteins (Matsuda and Garfinkel 2009). Antisense RNA could also act through the RNAi pathway to achieve EhLINE1 silencing, as seen with PIWI-interacting small RNAs (Iwasaki et al. 2015). In E. histolytica the RNAi pathway has been well characterized. It is mediated through small RNAs of 27 nt with a 5′-polyP structure (Zhang et al. 2008, 2011), and a 31 nt population that differs from the former in having 3′-4 non-templated As at the 3′-end (Zhang et al. 2020). Three homologues of sRNA-binding Argonaute protein (EhAgo) have also been reported (Zhang et al. 2019). These sRNAs have been shown to mediate long-term transcriptional gene silencing. It is possible that the antisense transcripts from EhLINE1 RT region could contribute to silencing through the sRNA pathway. Interestingly, the main genomic targets of both 27 nt and 31 nt sRNA populations seem to be the genomic EhLINEs and a set of 226 other gene loci (Zhang et al. 2020). It is not known whether these sRNAs are derived from the entire EhLINE sequence or mainly from the RT region. Further work on the mechanism of sRNA-mediated gene silencing in E. histolytica may reveal how these RNAs could contribute to EhLINE silencing, and whether a possible connection exists between the sRNAs and the 1.5 kb RT-antisense long noncoding RNA.

Possible influence of LINEs/SINEs on host gene expression and virulence

LINEs and SINEs are known to influence the expression of neighboring genes by a variety of mechanisms, for example by providing alternative promoters, splicing and polyadenylation sites and by heterochromatinization (Chow et al. 2010; Faulkner and Carninci 2009; Lev-Maor et al. 2003; Ostertag and Kazazian 2001). In parasites, the non-LTR retrotransposon, LmSIDER, of Leishmania major has a unique effect on parasite gene expression. LmSIDERs are found almost exclusively within the 3′-UTRs of L. major mRNAs. Interestingly, LmSIDER2-containing mRNAs are generally expressed at lower levels compared to the non-LmSIDER2 mRNAs, and LmSIDER2 is thought to act as mRNA instability element (Bringaud et al. 2007, 2008). In Trypanosomatid genomes, large directional polycistronic clusters, transcribed by RNA polymerase II, are separated by Strand Switch Regions (SSR). Transcription is mainly initiated at SSRs located between diverging clusters (Martinez-Calvillo et al. 2004). The location of non-LTR retrotransposon sequences in these genomes appears to be strongly biased within SSRs. In T. cruzi it has been suggested that the promoter Pr77 of L1Tc located at SSRs could participate in the transcription of adjacent genes and polycistrions, and that dispersed copies of L1Tc could prevent the decay of transcription level of distal genes in a cluster (Macias et al. 2016). TEs like INGI/RIME are inserted very frequently in the vicinity of VSG genes, which are responsible for antigenic variation. These TEs could have a possible role in the evolution of VSG gene repertoires by promoting recombination (Kimmel et al. 1987).

In Entamoeba, the role, if any, of TEs in influencing virulence potential is unknown. This could be studied by comparing E. histolytica and E. dispar which are very closely related, but while E. histolytica can invade tissues and cause disease, E. dispar is completely non pathogenic. One could expect differences at two levels. Firstly, the type and sequence of TEs could be very divergent in the two species; alternatively, TEs could be conserved in the two species, but their genomic location could be different. In both cases, TEs could be expected to mediate important phenotypic differences. Genome sequence analysis showed no major differences in the TE composition of E. histolytica and E. dispar, and LINEs and SINEs constituted a significant portion of the genome in both species. However, if some copies had moved to different genomic locations after the divergence of the two species, it could contribute significantly to the phenotypic differences as it is well known that LINE and SINE insertion can affect gene expression in a variety of ways. Investigations with
a limited number of genomic loci showed that the sites occupied by EhSINE1 in the *E. histolytica* genome were empty at homologous regions in *E. dispar* (Willhoef et al. 2002) and conversely the sites occupied by EdSINE in the *E. dispar* genome were empty in *E. histolytica* (Shire and Ackers 2007). Subsequently, a genome-wide comparison of the location of EhSINE in the *E. histolytica* and *E. dispar* genomes showed that only about 20% of syntenic sites were occupied by SINE1 in both species (Kumari et al. 2011). The absence of SINE1 in > 80% of syntenic loci in *E. histolytica* and *E. dispar* could result in differential expression of genes at these loci, which could be a source of phenotypic differences. The differential distribution of SINEs in the two genomes does not seem to be guided by the LINE-encoded endonuclease, as the EdLINE1 enzyme had the same kinetic properties and nicking site specificity as the enzyme coded by EhLINE1 (Kumari et al. 2011). It is not clear whether SINE expansion took place after the divergence of the two species, or SINEs occupied identical sites in the genome of the common ancestor of *E. histolytica* and *E. dispar* and were subsequently lost from many locations. Within *E. histolytica*, several lab-grown strains and clinical isolates were tested for EhSINE1 and EhSINE2 polymorphism. Seventeen polymorphic loci were identified where a EhSINE1/ EhSINE2 sequence was missing from the corresponding locus of one or more strains (Kumari et al. 2013). Some of these loci were associated with geographical location of *E. histolytica* strains (Sharma et al. 2017). While these data point to a possible association of LINE/SINE insertion polymorphism with virulence properties, much more work is needed to directly demonstrate such an effect.

In an analysis of the genes associated most frequently with TEs in *E. histolytica* (Lorenzi et al. 2008), the top three were hypothetical proteins. Important known protein families located close to TEs included gal/galNAc lectin, hsp70, BspA-like surface protein family and AIG family associated with resistance to bacteria. Of 28 transcribed EhLINE1 copies analyzed by RNA-Seq, only 3 were flanked by genes that were highly expressed, while the remaining were flanked by medium or low-expressing genes (Kaur et al. 2021). Further analysis of a similar nature with *E. dispar* genome could give interesting information on the possible contribution of TEs in regulating the expression of pathogenesis-related genes.

**Conclusion**

The study of LINEs in an early-branching eukaryote like *E. histolytica* serves to highlight the diversity of these retrotransposons in living systems. Phylogenetic analysis based on RT and EN sequences had placed EhLINEs in the ancient R2 group of retrotransposons that contain a single ORF and encode a REL-endonuclease domain. However, EhLINEs contain two ORFs, which is considered a feature of later evolving LINEs that acquired apurinic endonuclease domain, along with a second ORF (ORF1) 5′-of the RT-encoding ORF2. EhLINE1 appears to be exceptional in that it has acquired the ORF1 in a lineage that still contains the REL-endonuclease, indicating that the 5′-part of LINEs may, in fact, have evolved independently. As more LINEs are analyzed from a larger variety of organisms one may encounter other examples similar to EhLINEs.

TEs have co-evolved with their hosts so as to efficiently propagate themselves while causing minimal damage to the host genome. EhLINEs/SINEs use a variety of strategies in this regard. Although they are found commonly in gene-rich regions, they avoid inserting within genes. This selection of “safe” insertion sites seems to be guided, partly, by unique features of DNA structure upstream of the insertion site. Further, the retrotransposition activity of EhLINE1 is kept in check by the accumulation of point mutations, deletions and insertions in most copies; and by transcriptional silencing.

The novel transcription pattern of EhLINE1 seems to be well designed to limit retrotransposition by the near absence of full-length EhLINE1 transcripts, and by massive antisense transcription of the RT domain. The 1.5 kb antisense RT transcript could be involved in attenuating translation of the RT domain, as ORF2p was undetectable in *E. histolytica* cells, while ORF1p, for which there were no antisense transcripts, was expressed constitutively. In addition, it is possible that the antisense RT transcript, a long noncoding RNA, could have regulatory functions other than in retrotransposition. Similarly, the constitutively expressed ORF1p in cells not engaged in active retrotransposition leaves open the question of whether this polypeptide could have roles in cellular physiology other than those in retrotransposition. If so, these transposable elements could be essential for host fitness, and not merely selfish entities engaged in their own propagation.

**Acknowledgements** SB acknowledges financial support from the Indian National Science Academy. Parts of Fig. 1 were reproduced from Lorenzi et al. (2008), published by Springer Nature, under the terms of the creative commons license https://creativecommons.org/licenses/by/2.0.

**Author contributions** DK, MA and SB were involved in drafting the initial manuscript. All authors contributed to the final manuscript, read, and approved the submitted version.

**Declarations**

**Conflict of interest** The authors declare no conflicts of interest.
References

Agrahari M, Gaurav AK, Bhattacharya A, Bhattacharya S (2017) Cytosine DNA methylation at promoter of non-LTR retrotransposons and heat shock protein gene (HSP70) of Entamoeba histolytica and lack of correlation with transcription status. Mol Biochem Parasitol. https://doi.org/10.1016/j.molbiopara.2017.01.001

Aksoy S, Williams S, Chang S, Richards FF (1990) SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. Nucleic Acids Res. https://doi.org/10.1093/nar/18.4.785

Alexandrova EA, Olovnikov IA, Malakhova GV, Zabolotneva AA, Suntsova MV, Dmitriev SE, Buzdin AA (2012) Sense transcripts originated from an internal part of the human retrotransposon LINE-1 5′ UTR. Gene. https://doi.org/10.1016/j.gene.2012.09.026

Alish RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV (2006) Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. https://doi.org/10.1101/gad.1380406

Arkhipova IR, Morrison HG (2001) Three retrotransposon families in the genome of Giardia lamblia: two telomeric, one dead. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.231494798

Bagchi A, Bhattacharya A, Bhattacharya S (1999) Lack of a chromosomal copy of the circular RNA plasmid of Entamoeba histolytica. Int J Parasitol 29(11):1775–1783. https://doi.org/10.1016/s0020-7519(99)00125-3

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

Bakre AA, Rawal K, Ramaswamy R, Bhattacharya A, Bhattacharya S, Bakre AA (2005) Mobile genetic elements in protozoan parasites. J Genet. https://doi.org/10.1007/BF02715903

Bibilbo A, Eickbush TH (2002) The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. J Mol Biol. https://doi.org/10.1006/jmbi.2001.5369

Bibilbo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J Biol Chem. https://doi.org/10.1074/jbc.M310450200

Boeke JD (1997) Lines and Alus—the polyA connection. Nat Genet. https://doi.org/10.1038/ng0597-6

Boeke JD, Corces VG (1989) Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43:403–434. https://doi.org/10.1146/annurev.mi.43.100189.002155

Bringaud F, Bitezau N, Melville SE, Hez S, El-Sayed NM, Leech V, Berriman M, Hall N, Doneh JL, Balitz T (2002) A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of Trypanosoma brucei. Eukaryot Cell. https://doi.org/10.1128/EC.1.1.137-151.2002

Bringaud F, Müller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NMA, Papadopoulos B, Ghedin E (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania. PLoS Pathog. https://doi.org/10.1371/journal.ppat.0030136

Bringaud F, Ghedin E, El-Sayed NMA, Papadopoulos B (2008) Role of transposable elements in trypanosomatids. Microbes Infect. https://doi.org/10.1016/j.micinf.2008.02.009

Brouha B, Schustak J, Badge RM, Lutz-Frigge S, Farley AH, Moran JV, Kazazian HH (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.0831042100

Burke WD, Malik HS, Rich SM, Eickbush TH (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. Mol Biol Evol 19(5):619–630. https://doi.org/10.1093/oxfordjournals.molbev.a004121

Callahan KE, Hickman AB, Jones CE, Ghirlando R, Furano AV (2012) Polymerization and nuclear acid-binding properties of human L1 ORF1 protein. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr728

Chow JC, Ciaudol Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Ayer P, Wutz A, Barillot E, Greally JM, Voinnet O, Heard E (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell. https://doi.org/10.1016/j.cell.2010.04.042

Christensen SM, Eickbush TH (2005) R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. Mol Cell Biol. https://doi.org/10.1128/mcb.25.15.6617-6628.2005

Christensen SM, Bibilbo A, Eickbush TH (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPTR) reaction. Nucleic Acids Res. https://doi.org/10.1093/nar/gki957

Cost F, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. EMBO J. https://doi.org/10.1093/emboj/cdf392

Craik NL (2002) Mobile DNA: an introduction. In: Craig NL, Craig R, Gellert M, Lambertz AM (eds) Mobile DNA II. American Society for Microbiology, Washington, D.C., pp 3–11

Crichton JH, Dunican DS, MacLennan M, Meehan RR, Adams IR (2014) Defending the genome from the enemy within: Mechanisms of retrotransposon suppression in the mouse germline. Cell Mol Life Sci. https://doi.org/10.1007/s00018-013-1468-0

Criscione SW, Theodosakis N, Micevic G, Cornish TC, Burns KH, Neretti N, Rodic N (2016) Genome-wide characterization of human L1 antisense promoter-driven transcripts. BMC Genom. https://doi.org/10.1186/s12864-016-2800-5

Cruz-Reyes J, ur-Rehman T, Spiece WM, ACKERS JP (1995) A novel transcribed repeat element from Entamoeba histolytica. Gene. https://doi.org/10.1016/0378-1119(95)00549-X

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab. https://doi.org/10.1006/mgme.1999.2864

Delviks-Frankenberg K, Galli A, Nikolaiotchi O, Mens H, Pathak VK, Hu WS (2011) Mechanisms and factors that influence high frequency retroviral recombination. Viruses. https://doi.org/10.3390/v3091650

Denli AM, Narvaiza I, Kerman BE, Pena M, Benner C, Marchetto MCN, Diedrich JK, Aslanian A, Ma J, Moresco JJ, Moore L, Hunter T, Saghatelian A, Gage FH (2015) Primate-specific ORF0 contributes to retrotransposon-mediated diabetes. Cell. https://doi.org/10.1016/j.cell.2015.09.025

Derr LK, Strathern JN (1993) A role for reverse transcripts in gene conversion. Nature. https://doi.org/10.1038/361170a0

Dewannieux M, Enault S, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nature Genet. https://doi.org/10.1038/ng1223

Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, Kopera HC, Athanikar JN, Hasnouei M, Bachetton A, Moran JV, Gilbert N (2010) Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet. https://doi.org/10.1371/journal.pgen.1001150

Eickbush TH, Malik HS (2014) Origins and evolution of retrotransposons. Mobile DNA II. https://doi.org/10.1128/9781555817954.ch49

Elbarbary RA, Lucas BA, Maquat LE (2016) Retrotransposons as regulators of gene expression. Science. https://doi.org/10.1126/science.aac7247
repetitive elements in three Entamoeba species. BMC Genom. https://doi.org/10.1186/1471-2164-9-595
Luan DD, Korman MH, Jakubczak JL, Eickbusch TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. https://doi.org/10.1016/0092-8674(93)90078-5
Macías F, López MC, Thomas MC (2016) The Trypanosomatid Pr77-hallmark contains a downstream core promoter element essential for transcription activity of the Trypanosoma cruzi L1Tc retrotransposon. BMC Genom. https://doi.org/10.1186/s12864-016-2427-6
Macías F, Alfonso-Lehmann R, López MC, Gómez I, Thomas MC (2018) Biology of Trypanosoma cruzi retrotransposons: from an enzymatic to a structural point of view. Curr Genom 19(2):110–118. https://doi.org/10.2174/1389202918666170815150738
Mandal PK, Bagchi A, Bhattacharya A, Bhattacharya S (2004) An Entamoeba histolytica line/sine pair inserts at common target sites cleaved by the restriction enzyme-like line-encoded endonuclease. Eukaryot Cell. https://doi.org/10.1128/EC.3.1.170-179.2004
Mandal PK, Rawal K, Ramaswamy R, Bhattacharya A, Bhattacharya S (2006) Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to Entamoeba histolytica. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl710
Martin SL, Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol. https://doi.org/10.1128/mcb.21.2.467-475.2001
Martin F, Marañón C, Olivares M, Alonso C, López MC (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. https://doi.org/10.1006/jmbi.1994.0121
Martin SL, Crucenau M, Branciforte D, Li PWL, Kwok SC, Hodges RS, Williams MC (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol. https://doi.org/10.1016/j.jmb.2005.03.003
Martínez-Calvillo S, Nguyen D, Stuart K, Myler PJ (2004) Transcription initiation and termination on Leishmania major chromosome 3. Eukaryot Cell. https://doi.org/10.1128/EC.3.2.506-517.2004
Matsuda E, Garfinkel DJ (2009) Posttranslational interference of Ty1 retrotransposition by antisense RNA. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.0908305106
Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y (1992) Identification of an internal cis-element essential for the human l transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res. https://doi.org/10.1093/nar/20.12.3139
Mittal V, Bhattacharya A, Bhattacharya S (1994) Isolation and characterization of a species-specific multicopy DNA sequence from Entamoeba histolytica. Parasitology. https://doi.org/10.1017/S0031182000076071
Mizrokh L, Georgieva SG, Ilyin YV (1988) Jockey, a mobile drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. Cell. https://doi.org/10.1016/0092-8674(88)80013-8
Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87(5):917–927. https://doi.org/10.1016/s0092-8674(00)81998-4
Nakamura M, Okada N, Kajikawa M (2012) Self-interaction, nucleic acid binding, and nucleic acid chaperone activities are unexpectedly retained in the unique ORF1p of zebrfish LINE. Mol Cell Biol. https://doi.org/10.1128/mcb.06162-11
Okada N, Hamada M (1997) The 3’ ends of tRNA-derived SINEs originated from the 3’ ends of LINEs: a new example from the bovine genome. J Mol Evol. https://doi.org/10.1007/PL00000058
Okada N, Hamada M, Ogawa R, Ohshima K (1997) SINEs and LINEs share common 3’ sequences: a review. Gene. https://doi.org/10.1016/S0378-1119(97)00409-5
Ostertag EM, Kazazian HH Jr (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet. https://doi.org/10.1146/annurev.genet.35.102401.091032
Perepelitsa-Belancio V, Deininger P (2003) RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet. https://doi.org/10.1038/ng1269
Pritham EJ, Feschotte C, Wessler SR (2005) Unexpected diversity and differential success of DNA transposons in four species of Entamoeba protozoans. Mol Biol Evol. https://doi.org/10.1093/molbev/msi169
Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG (2012) The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res 40(4):1666–1683. https://doi.org/10.1093/nar/gkr863
Ralston KS, Petri WA (2011) The ways of a killer: how does Entamoeba histolytica elicit host cell death? Essays Biochem. https://doi.org/10.1042/BSE0510193
Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. https://doi.org/10.1038/nature05918
Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV (2015) The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Mobile DNA III. https://doi.org/10.1128/9781555819217.ch51
Sharma R, Bagchi A, Bhattacharya A, Bhattacharya S (2001) Characterization of a retrotransposon-like element from Entamoeba histolytica. Mol Biochem Parasitol. https://doi.org/10.1016/S0166-6851(01)00300-0
Sharma S, Banyal N, Singh M, Mandal AK, Bhattacharya S, Paul J (2017) SINE polymorphism reveals distinct strains of Entamoeba histolytica from North India. Exp Parasitol. https://doi.org/10.1016/j.exppara.2017.01.007
Shire AM, Ackers JP (2007) SINE elements of Entamoeba dispar. Mol Biochem Parasitol. https://doi.org/10.1016/j.molbiopara.2006.11.010
Silberman JD, Clark CG, Diamond LS, Sogin ML (1999) Phylogeny of the genera Entamoeba and Endolimax as deduced from small-subunit ribosomal RNA sequences. Mol Biol Evol 16(12):1740–1751. https://doi.org/10.1093/molbev/sme169
Soifer HS, Rossi JJ (2006) Small interfering RNAs to the rescue: blocking L1 retrotransposition. Nat Struct Mol Biol. https://doi.org/10.1038/nsmb0906-758
Speek M (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol. https://doi.org/10.1128/MCB.21.6.1973-1985.2001
Stenglein MD, Harris RS (2006) APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J Biol Chem. https://doi.org/10.1074/jbc.M602367200
Swergold GD (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol. https://doi.org/10.1128/MCB.10.12.6718.2000
Thomas MC, Macías F, Alonso C, López MC (2010) The biology and evolution of transposable elements in parasites. Trends Parasitol. https://doi.org/10.1016/j.pt.2010.04.001
van Dellen K, Field J, Wang Z, Lofts B, Samuelson J (2002) LINEs and SINE-like elements of the protist Entamoeba histolytica. Gene. https://doi.org/10.1016/S0378-1119(02)00917-4
Vázquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A, Levin MJ (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. https://doi.org/10.1073/pnas.050578397

Villanueva MS, Williams SP, Beard CB, Richards FF, 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. https://doi.org/10.1128/mcb.11.12.6139

Willhoeft U, Buß H, 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. https://doi.org/10.1016/S1434-4610(99)70009-X

Willhoeft U, Buß H, Tannich E (2002) The abundant polyadenylated transcript 2 DNA sequence of the pathogenic protozoan parasite Entamoeba histolytica represents a nonautonomous non-long-terminal-repeat retrotransposon-like element which is absent in the closely related nonpathogenic species Entamoeba dispar. Infect Immun. https://doi.org/10.1128/IAI.70.12.6798-6804.2002

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. https://doi.org/10.1002/j.1460-2075.1990.tb07536.x

Yadav VP, Mandal PK, Rao DN, Bhattacharya S (2009) Characterization of the restriction enzyme-like endonuclease encoded by the Entamoeba histolytica non-long terminal repeat retrotransposon EhLINE1. FEBS J. https://doi.org/10.1111/j.1742-4658.2009.07419.x

Yadav VP, Mandal PK, Bhattacharya A, Bhattacharya S (2012) Recombinant SINEs are formed at high frequency during induced retrotransposition in vivo. Nat Commun. https://doi.org/10.1038/ncomms1855

Yang J, Malik HS, Eickbush TH (1999) Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.96.14.7847

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. https://doi.org/10.1016/S0168-9525(97)01181-5

Yu F, Zingler N, Schumann G, Strätling WH (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. Nucleic Acids Res. https://doi.org/10.1093/nar/29.21.4493

Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U (2008) Small RNAs with 5′-polyphosphate termini associate with a piwi-related protein and regulate gene expression in the single-celled eukaryote Entamoeba histolytica. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1000219

Zhang H, Alamini H, Tran V, Singh U (2011) Nucleus-localized antisense small RNAs with 5′-polyphosphate termini regulate long term transcriptional gene silencing in Entamoeba histolytica G3 strain. J Biol Chem. https://doi.org/10.1074/jbc.M111.278184

Zhang H, Tran V, Manna D, Ehrenkaufer G, Singh U (2019) Functional characterization of Entamoeba histolytica argonaute proteins reveals a repetitive DR-rich motif region that controls nuclear localization. Msphere. https://doi.org/10.1128/msphere.00580-19

Zhang H, Ehrenkaufer GM, Hall N, Singh U (2020) Identification of oligo-adenylated small RNAs in the parasite Entamoeba and a potential role for small RNA control. BMC Genom. https://doi.org/10.1186/s12864-020-07275-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.