Plasmid 114 (2021) 102560

Contents lists available at ScienceDirect

Plasmid

journal homepage: www.elsevier.com/locate/yplas

Transcriptomic analysis of *Entamoeba histolytica* reveals domain-specific sense strand expression of LINE-encoded ORFs with massive antisense expression of RT domain

Devinder Kaur [a,1,2], Mridula Agrahari [a,1], Shashi Shekhar Singh [a,3], Prabhat Kumar Mandal [b], Alok Bhattacharya [c], Sudha Bhattacharya [a,4,*]

[a] School of Environmental Sciences, Jawaharlal Nehru University, India
[b] Department of Biotechnology, IIT Roorkee, India
[c] Ashoka University, Sonepat, India

ARTICLE INFO

Keywords:
EhLINE1
Antisense RNA
LINE transcriptome
Truncated transcripts
RNA-Seq
E. histolytica
Internal deletion
Read-through transcription

ABSTRACT

LINEs are retrotransposable elements found in diverse organisms. Their activity is kept in check by several mechanisms, including transcriptional silencing. Here we have analyzed the transcription status of LINE1 copies in the early-branching parasitic protist *Entamoeba histolytica*. Full-length EhLINE1 encodes ORF1, and ORF2 with reverse transcriptase (RT) and endonuclease (EN) domains. RNA-Seq analysis of EhLINE1 copies (both truncated and full-length) showed unique features. Firstly, although 20/41 transcribed copies were full-length, we failed to detect any full-length transcripts. Rather, sense-strand transcripts mapped to the functional domains- ORF1, RT and EN. Secondly, there was strong antisense transcription specifically from RT domain. No antisense transcripts were seen from ORF1. Antisense RT transcripts did not encode known functional peptides. They could possibly be involved in attenuating translation of RT domain, as we failed to detect ORF2p, whereas ORF1p was detectable. Lack of full-length transcripts and strong antisense RT expression may serve to limit EhLINE1 retrotransposition.

1. Introduction

Long Interspersed Nuclear Elements (LINEs) are a class of non-long terminal repeat (non-LTR)-containing retrotransposon, that are ubiquitously present in most genomes across the phylogenetic spectrum ranging from unicellular protists to plants and mammals (Eickbush and Malik, 2014). Yet their origin and sustenance remain mysterious, and their complex relationship with the host genome and with cellular physiology is not clearly understood (Goodier, 2016; Han, 2010; Kazazian, 2004). A common feature of these elements is that although they exist in a large number of copies in extant genomes, they are generally maintained in a transcriptionally silent state with only a few copies being active (Huang et al., 2012; Ostertag and Kazazian Jr, 2001; Sassaman et al., 1997). In addition, most copies contain multiple mutations, including 5′- and 3′- truncations, point mutations and large deletions, which render them non-functional. These strategies limit their active insertion in genes, which could be lethal for the host genome (Deininger and Batzer, 1999).

LINEs show a fairly conserved organization of functional features. The prototype LINE may be represented by the human LINE-1 (L1),

Abbreviations: RSEM, RNA-Seq by expectation-maximization; TPM, Transcripts per million; LINE, Long Interspersed Nuclear Element; UTR, Untranslated region; ORF, Open Reading Frame; RT, Reverse Transcriptase domain; EN, Endonuclease domain; CIs, credibility intervals; PME, Posterior Mean Estimate; ML, Maximum Likelihood; MSA, Multiple Sequence Alignment; MAST, Motif alignment and search tool; MEME, Multiple Expectation maximizations for Motif Elicitation; FIMO, Find Individual Motif Occurrences.

* Corresponding author at: School of Environmental Sciences, Jawaharlal Nehru University, India.
E-mail addresses: pkm31fbt@iitr.ac.in (P.K. Mandal), alok.bhattacharya@gmail.com (A. Bhattacharya), sbjnu110@gmail.com, sudha.bhattacharya@ashoka.edu.in (S. Bhattacharya).
[1] Joint First authors.
[2] Present address: Central University of Punjab, Bhatinda.
[3] Present address: Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA
[4] Present address: Ashoka University, Sonepat, India

https://doi.org/10.1016/j.plasmid.2021.102560
Received 21 September 2020; Received in revised form 29 December 2020; Accepted 31 December 2020
Available online 20 January 2021

which has been most extensively studied. Human L1 occupies around 17% of the genome (Lander et al., 2001) with an estimated 500,000 copies, of which there are about 7000 full-length copies, and only 80–100 are retrotransposition competent (Brouha et al., 2003; Rangwala et al., 2009). Full-length L1 is around 6.0 kb in length, contains 5′-UTR with an internal sense as well as antisense promoter, 2 open reading frames (ORFs) designated as ORF1 and ORF2, and 3′-UTR which ends with variable length poly(A) sequence (Scott et al., 1987). Transcription from the internal promoter in a LINE element could give a full-length bicistronic transcript whose translation would provide the two polypeptides (ORF1p and ORF2p) required for retrotransposition. However, early studies revealed the complex nature of LINE transcription. Experiments with a variety of systems, including Drosophila I element (Chaboissier et al., 1990), mammalian LINE elements from human and rodent cell lines and various tissues (Benihoud et al., 2002; Dudley, 1987; Martin, 1991; Packer et al., 1993; Perepelitsa-Belancio and Deininger, 2003) showed that LINE element is transcribed into a heterogeneous population of RNAs (ranging in size from 0.2 kb to full-length). These were mostly sense strand transcripts, although some antisense transcripts were also detectable (Benihoud et al., 2002; Speek, 2001).

Due to the very large copy number of LINEs and the high sequence similarity between copies, it has been difficult to determine the expression of individual endogenous copies and study the transcription pattern of each copy. A considerable amount of human L1 transcription does not originate from its internal promoter as a large number of L1s are inserted in introns and are co-transcribed within genes. In these instances, different L1 loci are expressed in different tissues without the use of L1 promoter (P. Deininger et al., 2017; Kaul et al., 2020). Transcription from the L1 promoter is high in germ cells, at early stage of embryonic development, in certain tumors, and in neuronal progenitor cells (Belancio et al., 2010; Coufal et al., 2009; Faulkner et al., 2009; Wissing et al., 2012). This transcription is contributed by the L1 HS-Ta subfamily which is considered to be the youngest, and consists of currently transcriptionally active members. On an average there are an estimated 652 (±68) L1HS-Ta copies per cell line. The transcriptional activity of individual L1HS-Ta copies was measured (Philippe et al., 2016). Only a small subset of L1HS-Ta loci contributed to the bulk of L1 expression, and the genomic environment in which the L1 copy was inserted had an important role in determining its expression. Thus, the transcriptional status of individual LINE copies in a genome is variable, and may have important functional implications for the host organism.

The transcriptional behaviour of individual LINE copies could be simpler to analyze in an early-branching unicellular protist in which the copy number of these elements is much smaller than human, and the elements are mostly inserted in intergenic regions due to the paucity of introns in protein-coding genes. We have been studying the human parasitic protist, *Entamoeba histolytica*, which has three classes of LINEs: EhLINE1, EhLINE2 and EhLINE3. Among these, EhLINE1, present in an estimated 742 copies is the most abundant (Bakre et al., 2005; Lorenzi et al., 2008). Knowledge of the transcription pattern of genomic EhLINE1 copies would be helpful to decipher any possible correlation between differential expression of virulence genes in *E. histolytica* strains of low and high virulence with EhLINE1 insertion polymorphism. In earlier studies with EhSINE1 (the nonautonomous partner of EhLINE1) we have shown that geographical isolates of *E. histolytica* exhibit marked SINE insertion polymorphism (Kumari et al., 2013; Sharma et al., 2017). Hence, we are interested to investigate the transcriptional behaviour of individual EhLINE1 copies to reveal the basic features of EhLINE1 transcription which are so far unknown, and to use this information for future analysis of differential gene expression in *E. histolytica* isolates. The influence of non-LTR retrotransposons on parasite gene expression has been reported in the protozoan parasite *Leishmania major* which contains non-LTR retrotransposons called LmSIDERs. These 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). We expect that information about transcriptional activity of individual EhLINE1 copies would be useful to understand the influence of these elements on *E. histolytica* gene expression. In addition, this study with an evolutionarily distant organism would provide an interesting comparison with the known transcription patterns of human L1.

Full-length EhLINE1 is ~4.8 Kb in length, with two non-overlapping ORFs: ORF1 of 1.5 Kb and ORF2 of ~3 Kb separated by a spacer region of ~440 bp. The domain structures of EhLINE1-encoded ORFs have similarities with human L1-encoded ORFs. EhLINE1 ORF1-encoded protein contains an RNA recognition motif with single stranded nucleic acid binding activity whereas ORF2 codes for protein with reverse transcriptase (RT) and endonuclease (EN) activities (Gaurav et al., 2017; Mandal et al., 2004; Yadav et al., 2009). Unlike human L1, the endonuclease encoded by EhLINEs is restriction enzyme-like, as seen in elements of the R4 clade to which EhLINEs belong (Eickbush and Malik, 2014; Yang et al., 1999). This enzyme has also been shown to have sequence and structure homology with archaeal Holliday junction resolvases, and this activity is utilized to achieve second-strand DNA synthesis during LINE integration (Khadgi et al., 2019). Most EhLINE1 copies are truncated or mutated, and *E. histolytica* cells normally do not express detectable levels of ORF2, although ORF1 polypeptide is constitutively expressed. Upon ectopic overexpression of ORF2, *E. histolytica* cells could be made retrotransposition-competent (Yadav et al., 2012). Here we have used RNA-Seq to determine the transcription status of individual EhLINE1 copies (both truncated and full-length). We find that sense transcripts primarily map to the functional domains, namely ORF1, RT and EN, with the absence of full-length transcripts. Antisense transcripts, which exceed sense transcripts in number, are almost exclusively derived from the RT domain. To our knowledge, this novel transcription pattern has not been reported for other LINEs. It may be designed to limit retrotransposition both by restricting the number of full-length transcripts and by attenuating RT expression.

2. Materials and methods

2.1. Cell culture and growth conditions

Trophozoites of *E. histolytica* strain HM-1:IMSS were axenically maintained in TYI-S-33 medium supplemented with 15% adult bovine serum, 1× Diamond's vitamin mix and antibiotic (125 μl of 250 units/ml benzyl penicillin and 0.25 mg/ml streptomycin per 90 ml of medium) at 35.5 °C (Diamond et al., 1978) in normal condition and for 60 min at 42 °C during heat stress condition. As *E. histolytica* is microaerophilic, oxidative stress was induced by providing aeration to the cells.

2.2. Isolation of total RNA from E. histolytica trophozoites

One million trophozoites (50 ml culture) growing in log phase were harvested at 600 x g for 5 min at 4 °C. The cell pellet was washed with ice chilled PBS #8 (0.37% K2HPO4, 0.11% KH2PO4 and 0.95% NaCl, pH 7.2) and resuspended in 1 ml of Trizol reagent (Invitrogen) followed by lysis through repeated pipetting. RNA isolation was carried out according to the manufacturer's protocol (Invitrogen). Briefly, the lysed cells were incubated at room temperature for 10–15 min. 200 μl of chloroform was added and the mixture was shaken vigorously for 15–30 s followed by incubation for 10–15 min. The tubes were centrifuged at 12000 x*g* for 15 min at 4 °C for complete phase separation. The upper aqueous phase containing RNA was transferred to a fresh tube and RNA was precipitated with 500 μl of isopropanol at room temperature for 10 min. RNA pellet was collected by centrifugation at 12000 x*g* for 10 min at 4 °C. Pellet was washed with 1 ml of chilled 70% ethanol in DEPC treated water (freshly prepared) at 7500 x g for 5 min at 4 °C. The pellet was dried at 37 °C for 15 min and resuspended in 50 μl of DEPC-treated water, aliquoted and stored at −80 °C.

2.3. Isolation of poly(A) + RNA

Poly(A) + RNA was purified from total RNA using poly(A) tract mRNA isolation system III from Promega (Z5300) as per manufacturer protocol. Briefly, total RNA was incubated in a sterile tube for 10 min at 65 °C. A biotinylated oligo(dT) probe and SSC was added to the RNA and incubated at room temperature for 10 min to hybridize with the 3′-poly (A) + region. The hybrids were added to the washed streptavidin-coupled paramagnetic particles, captured using a magnetic separation stand and washed at high stringency with the provided buffer. The purified poly(A) + RNA was eluted from the solid phase by the addition of provided ribonuclease-free, deionized water.

2.4. Transcriptome analysis

This was done essentially as previously described (Naiyer et al., 2019). The non-stranded RNA-Seq dataset used in this work have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE151975 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151975). Total RNA, from three biological replicates, was used for selection of poly(A) + RNA and library preparation was done after oligo(dT) selection. RNA-Seq libraries were generated by performing RNA fragmentation, random hexamer primed cDNA synthesis, linker ligation and PCR enrichment. These libraries were then subjected to paired-end sequencing on the Illumina HiSeq2500 (v3 Chemistry) platform. A total of 222,636,746 paired-end reads of size 100 bp were generated from the 3 biological replicates.

To align the high quality reads to *E. histolytica* (HM-1:IMSS) genome, accession number: AAFB00000000.2, the gene model was downloaded from AmoebaDB (http://amoebadb.org/common/downloads/release-27/EhistolyticaHM1IMSS/gff/data/). The alignment was performed using Tophat program (version 2.0.11) with default parameters. The RSEM program (version 1.3.0) was used using default values for estimating expression of the EhLINE1 elements using EhLINE1 transcripts as reference sequences (Li and Dewey, 2011). The pre-processed reads were aligned to 742 LINE1 fasta sequences of *E. histolytica* (HM1:IMSS) genome, accession number: AAFB00000000.2, which were extracted from AmoebaDB sequence retrieval section using "Retrieve Sequences By Genomic Sequence IDs". EhLINE1 copies with expected read count >10 in all 3 biological replicates were considered as expressed copies.

2.5. Strand specific expression analysis

Total RNA of *E. histolytica* was used for library preparation from two biological samples following protocol of Illumina TruSeq stranded total RNA sample preparation guide (Illumina, 2013) using the reagents provided in TruSeq Stranded Total RNA with Ribo-Zero™ Human/Mouse/Rat.

The libraries generated were then subjected to paired-end sequencing on the Illumina HiSeq2500 (v3 Chemistry) platform. Around 53 and 57 million reads of size 100 bp each were generated. After trimming adapter sequences (the parameter used for trimming was LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:75) and removing low quality reads using FastQC (Andrews, 2010) we aligned reads to *E. histolytica* genome. For strand specific expression study, the same RSEM pipeline was used including the option of "–strandedness reverse/forward" using default values.

2.6. Visualization of expression data

Wiggle plots and IGV viewer were used for visualizing the alignment files. Wiggle plots were generated using "rsem-plot-transcript-wiggles" which requires the sample name (expression file name) and transcript IDs. The output was a pdf file containing plots for the entire provided transcript IDs and sorted bam file of the sample. Further, for in-depth study, expression data was also visualized in IGV viewer. Of this sorted bam file (generated in wiggle plot step) we generated its index file of same name suffix with .bai using samtools (H. Li et al., 2009). Sorted bam file and respective bai file were used to visualize expression data in IGV viewer.

2.7. Sequence analysis of EhLINE1 genomic copies

All 742 genomic copies of EhLINE1 were categorized as full-length, truncated, and internal deletions by multiple sequence alignment using MAFFT (Katoh and Standley, 2013). Small sized copies were manually aligned to the full-length EhLINE1 (DS571192). We considered a copy full-length if it was within 5 nt at each end and did not have any large internal deletions. We considered a copy 5′ and 3′ truncated if it had truncation of more than 20 nt at either end of EhLINE1.

2.8. EhLINE1 transfectants

Full-length EhLINE1 (4.8 kb) was cloned in place of CAT in the tet-inducible vector pEhHYG-tetR-O-CAT (Hamann et al., 1997) at *Kpn*I and *Bam*HI sites, essentially as previously described (Yadav et al., 2012). The 5′- and 3′-actin and 5′-lectin sequences contain regulatory sequences from the *E. histolytica* genes to drive transcription. A sequence of 19 bp inserted between the TATA box and ATTCA initiator element in the lectin promoter acts as a TetR-operator. Transfection was carried out by electroporation and stable transfectants were maintained with 10 μg/ml of hygromycin B as previously described (Singh et al., 2018).

2.9. Northern blotting, hybridization, and probe preparation

RNA (20–30 μg) was denatured by incubating with 2× RNA loading dye at 65 °C for 15 min followed by snap chilling on ice. Samples were loaded on denaturing agarose gel containing 2.2 M formaldehyde and 1× MOPS buffer followed by electrophoresis at 4 V/cm. The gel was washed extensively with DEPC treated water to remove the formaldehyde and sequentially treated with denaturing (0.05 N NaOH and 1.5 M NaCl) and neutralizing solution (0.5 M Tris-HCl pH 7.5 and 1.5 M NaCl) for 20 min each, followed by 20 min equilibration in 20× SSC. The transfer membrane [nylon membrane (GeneScreen Plus; PerkinElmer)] was pre-equilibrated with 20× SSC and the blotting was done by passive transfer using standard protocols. After transfer, the RNA was UV cross-linked and blot was stained with methylene blue to check equal loading and to detect size of the molecular marker.

DecaLabel DNA Labeling Kit (Thermo Scientific #K0622) which is based on random priming method, was used for probe preparation according to manufacturer's protocol. Briefly, 50–100 ng of linear DNA along with Decanucleotide in 5× Reaction Buffer was denatured by heating in a boiling water-bath for 10 min followed by snap chilling on ice. To the tube containing denatured DNA, 3 μl of mixA, 30–50 μCi [α-32P] dATP and 5 U of Klenow enzyme were added and the reaction was incubated at 37 °C for 5 min followed by incubation with 4 ul of dNTP mix at 37 °C for 5 min. 1 μl of 0.5 M EDTA, pH 8.0 was added to stop the reaction. Unincorporated dNTPs were removed by ethanol precipitation in the presence of 50 μg of carrier DNA (salmon sperm DNA) and 2.5 M ammonium acetate, or by nucleotide removal kit (Qiagen).

RNA blots were first incubated in prehybridization solution (1% SDS and 1 M NaCl, 0.3–0.4 ml per square cm of the membrane) at 65 °C in hybridization bottles. After 3 h, heat-denatured radiolabeled probe (2 × 105 dpm/ml) and 100 μg/ml denatured salmon sperm DNA was added to the prehybridization mix and hybridization was carried out for 16 h at 65 °C. The membranes were washed sequentially twice with 2× SSC at RT for 5 min, twice with 1× SSC and 1% SDS at 65 °C for 30 min and finally twice with 0.1× SSC at RT for 30 min each to remove non-specifically bound probe.

For the radiolabeled strand specific DNA probes, region 1 (ORF1 full-

length) and region 2 (ORF2 B + C) were used as template for amplification with the primer set HJ67FP + EK39RP and BK49FP + DY32RP respectively. Purified template together with one primer from each primer set was used for linear PCR as described (Kimpton et al., 1993) with few modifications. Amplification reaction contained 1× Taq polymerase buffer, 200 μM each dA/G/TTP, and 5 μM dCTP, 50 μCi [α-P32] dCTP, 30 pmol respective primers, 10 ng/kb DNA template and 5 U of *Taq*DNA polymerase in a reaction volume of 50 μl. The linear PCR cycle comprised of an initial denaturation at 94 °C for 3 min followed by 40 cycles of denaturation at 94 °C for 30 s, annealing for 1 min at the Tm of the primer, extension at 72 °C for 90 s. The last extension step at 72 °C was done for an additional 10 min. Unincorporated dNTPs were removed by nucleotide removal kit (Qiagen).

2.10. Primers used

The nucleotide sequences of all primers used in this study are listed in Additional file 1.

2.11. RT-PCR analysis

For the RT-PCR analysis oligo(dT) primer was used for reverse transcription followed by PCR with ORF1 and RT-specific primer pairs. To minimize non-specific reverse transcription with oligo(dT) (since *E. histolytica* genome is highly A + T rich), a 45-mer primer was used and RT reaction was performed at high temperature at 50 °C for 1 h.

2.12. Luciferase reporter assay

The assay was performed as described (Shrimal et al., 2010). Briefly, stably transfected trophozoites, were washed in 1× PBS (pH 7.4) lysed in 200 μl of reporter lysis buffer (Promega) with the addition of protease inhibitor cocktail (Sigma) and were frozen overnight at −80 °C. Lysates were thawed on ice and pelleted to remove cellular debris. Before measuring the activity, samples were allowed to warm at room temperature, and assayed according to the manufacturer's instructions (Promega) using a Luminometer (Promega). Luciferase activity per μg of protein was calculated. Statistical comparisons were made using analysis of variance test and *t*-test. Experimental values were reported as the mean (±S.D.) for triplicate values. All calculations of statistical significance were made using the GraphPad InStat software package (GraphPad Prism 9).

3. Results

3.1. Organization of EhLINE1 copies in E. histolytica

In order to undertake the transcriptomic analysis of individual EhLINE1 copies we first determined their sequence organization. Genome-wide analysis of the 742 EhLINE1 copies had earlier shown that 88 were full-length while the rest were truncated (Lorenzi et al., 2008). We further identified the sequence features of each copy, including deletion breakpoints in truncated copies. The 742 copies spanned a wide range of sizes, from 42 bp to 4811 bp, with most copies in the range of 1–2 kb (Additional file 2: Fig. S1). Our analysis showed that only 61 copies were full-length (size range 4589 bp to 4811 bp), as opposed to the previous count of 88 full-length copies. This was because 27 of these copies had large internal deletions, although their 5′- and 3′-ends were intact. Details of the number of truncated EhLINE1 copies, with one or both ends truncated, are given in Additional File 2: Fig. S2. In copies with one or more internal deletions, the size of internal deletions ranged from 38 bp to 2.5 kb. Interestingly, the deleted sequences were flanked by direct repeats of size 5–34 nt, with sequence identity of 67–100% (Additional File 2: Figs. S3–S4). The presence of flanking direct repeats is significant as these are known to be targets of recombination, or replication slippage, leading to internal deletions (Pierce et al., 1991; Trinh and Sinden, 1993). Microhomology-mediated end joining for repair of DNA double strand breaks could also result in deletions (Seol et al., 2018; Sfeir and Symington, 2015). This analysis helped to comprehensively describe the sequence arrangement of all 742 EhLINE1 copies.

3.2. Identification of EhLINE1 transcripts by RNA-Seq

We performed RNA-Seq analysis to determine the expression status of all EhLINE1 copies, using poly(A)-enriched RNA of *E. histolytica*, grown under normal conditions, from three independent cultures. Details of the transcriptome analysis are presented in Additional File 2: Fig. S5. After trimming adapter sequences and removing low quality reads using trimmomatic-0.36, we got 67 to 71 million reads per sample (GEO Series accession number GSE151975). The GC content was 33% and the percentage of reads with ≥Q30 were 99.95% in all three biological replicates. On an average, ~93.70% of total high quality reads (208,484,400) aligned to *E. histolytica* (HM-1:IMSS) genome.

To map the reads on the 742 EhLINE1 copies we excluded the small copies of size <300 nt (136 copies) from further analysis. EhLINE1 copies with expected read count >10 in all biological samples were considered as expressed copies. We found that the most suitable method for analysis of our RNA-Seq data was RSEM, which has been used for RNA-Seq read mapping to TEs (Jin et al., 2015). RSEM is a generative probabilistic model, designed to address the issue of read mapping uncertainty and, therefore, to produce more accurate gene expression estimates (Li et al., 2010). RSEM was shown to provide improved accuracy with both mouse and maize. The improvement in accuracy was most striking for repetitive genomes, such as maize, which give rise to large fractions of multireads. It was used to generate leaf development transcriptome data in maize (Hughes et al., 2014), and is comparable with other TE-specific expression analysis tools (Jeong, H. H et al., 2018). Some of the more recent methods (Yang et al., 2019) could not be used as they need annotation files with assembled genomes available on UCSC genome browser and Repeatmasker file for repeats of interest. However, for our organism, *E. histolytica*, the annotation files with assembled genomes are not available on UCSC genome browser. Also, Repeatmasker would not work as the data for repeat elements of *E. histolytica* is not available in their back-end database. Thus, RSEM was the best available option. The reads were mapped to the whole genome and those mapping to EhLINE1 were extracted. Alternatively, reads were directly mapped to EhLINE1. Both approaches gave similar output.

Given the high sequence similarity of EhLINE1 copies, there were instances of uncertainty associated with assigning reads to individual copies. This was resolved by using credibility intervals (CIs). We used 95% CIs for the abundance estimates to examine uncertainty (Gupta et al., 2012; Li and Dewey, 2011). The posterior mean estimate (PME) values were used in lieu of maximum likelihood (ML) estimates as these values were generally contained within the 95% CIs. Thus, we used pme_TPM values for correlation calculation and further expression analysis. The Spearman correlation among three biological samples for EhLINE1 was $R \geq 0.84$ (Additional File 2: Fig. S6). Data showed that of the 606 copies of EhLINE1 analyzed, only 41 copies (6.7%) were transcriptionally active. Of these, 20 were full-length copies while 21 had internal deletions/end truncations (Fig. 1). Of the latter 21 copies, 10 had 5′ truncations, 3 had 3′ truncation, 3 had both ends truncated, and 5 had both ends intact (with internal deletions). Details of truncation status and sense/antisense transcription of all 41 expressed copies are given in Table 1, and they are further described in later sections.

Next, we mapped the reads from each expressed copy to determine whether the entire EhLINE1 sequence was transcribed. We found that reads mainly corresponded to three regions of EhLINE1: (1) 5′-end to 1517 bp; (2) 2440 to 3791 bp; (3) 3849 to 3′-end. These regions corresponded to ORF1, and to the RT and EN functional domains of ORF2 respectively (Fig. 2) (for description of RT and EN domains see (Mandal et al., 2004). Very few reads mapped between 1510 and 2430 bp which

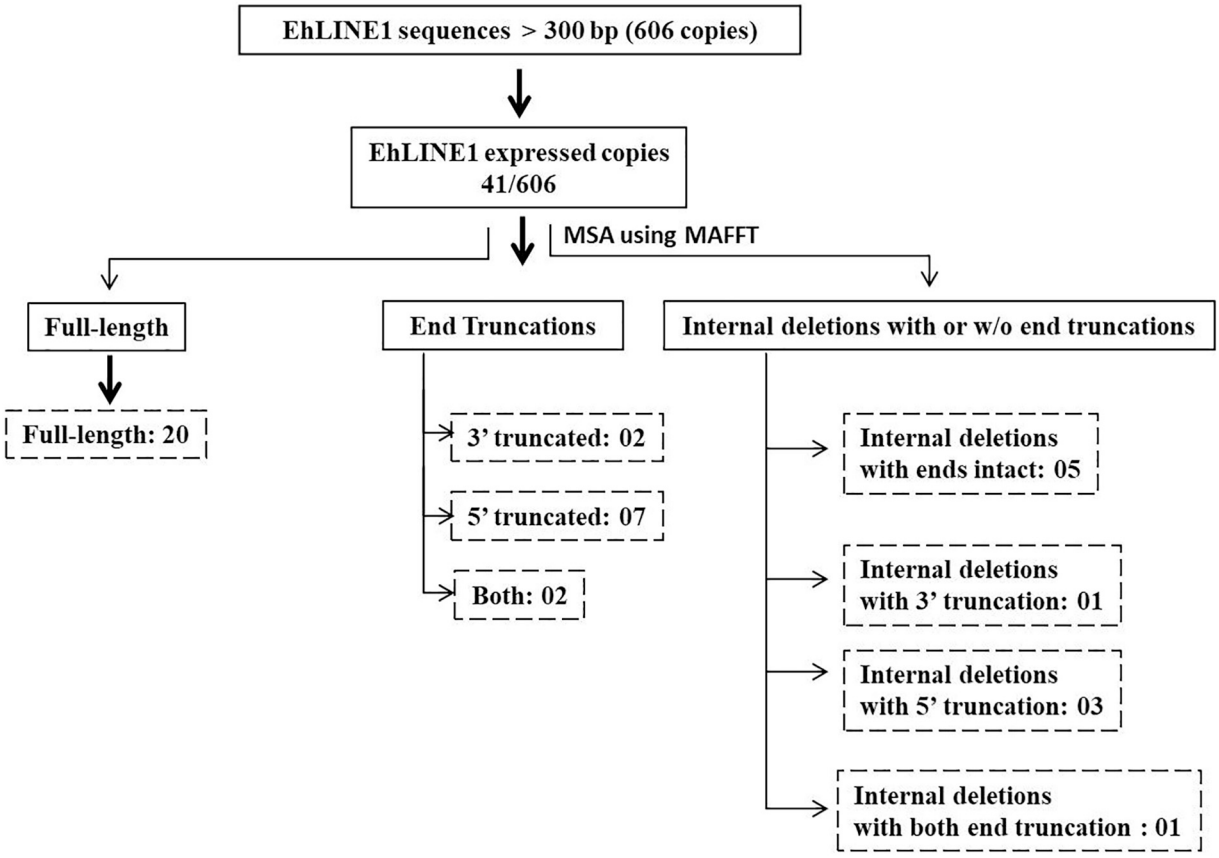

Fig. 1. Distribution of transcriptionally active genomic copies of EhLINE1.Of the total 742 EhLINE1 copies, only 606 were selected for analysis which were >300 bp.

included the regions between ORF1 and ORF2 (spacer) and the 5′-part of ORF2 that did not match with RT domain. Mapping of RNA-Seq reads indicated that intracellular, steady state EhLINE1 transcripts were truncated and corresponded to the functional protein domains.

The RNA-Seq data were validated by northern blot analysis using a panel of DNA probes spanning the entire length of EhLINE1 (Fig. 3). We divided the EhLINE1 into three regions where the maximum RNA-Seq reads mapped. These were region 1 (ORF1), region 2 (RT domain), region 3 (EN domain). Probes were designed from these three regions, and also from region 1–2 lying between ORF1 and RT where very few reads mapped. Northern data with these probes corroborated very well with RNA-Seq data (Fig. 3B). No 4.8 kb band corresponding to full-length EhLINE1 was visible with any of the probes. The ORF1 probe hybridized with a 1.5 kb band, and the expected transcript size from RNA-Seq reads was 1517 nt. The RT probes hybridized with a broad band slightly smaller than 1.5 kb, and the expected transcript size from RNA-Seq reads was 1351 nt. Region 1–2 probes failed to give any signal. To confirm that this was not due to poor labelling of probes, we performed dot blot hybridization with total genomic DNA (Fig. 3C), in which the probes gave bright signals. The EN probe did not give any signal in northern blot although RNA-seq reads were obtained from this region. It is possible that transcripts of this region were mainly short and heterogeneous in size and therefore not visible in northern analysis. We could detect transcripts from this region by RT-PCR (Fig. 3D). The absence of full-length (4.8 kb) RNA band in northern blots was not due to a technical problem of RNA prep, as we could see the 4.8 kb band in *E. histolytica* cells transfected with full-length EhLINE1 driven from a tetracycline-inducible promoter (Fig. 4). No 4.8 kb band was seen in these cells in the absence of tet.

The large number of RNA-Seq reads from RT region compared with other regions was due to massive antisense transcription, as described below.

3.3. Massive antisense transcription from RT region of EhLINE1

LINE elements are known to be transcribed in the antisense orientation, although the extent of sense transcription is generally greater. We looked at a strand-specific transcription of EhLINE1 by RNA-Seq. Two biological samples were analyzed for this study following the Illumina TruSeq stranded total RNA protocol. The Spearman correlation among two biological samples for EhLINE1 was $R \geq 0.76$ (Additional File 2: Fig. S7). After trimming adapter sequences and removing low quality reads using FastQC, we got 37,797,969 and 35,845,132 reads which showed 32% mapping rate to *E. histolytica* genome. The low mapping rate was due to presence of reads from the rRNA region of *E. histolytica* as confirmed by BLASTn analysis. On aligning the sequence reads, we found sense transcripts from 3 regions of EhLINE1 (ORF1, RT and EN) but antisense transcripts were coded almost exclusively from RT region, with a very small number of antisense reads coming from ORF1 and EN regions (Fig. 5). The antisense reads from RT region mapped to distinct 5′- and 3′-ends. We looked at the 41 expressed copies individually to score for copies expressed in sense or antisense orientation, keeping read count ≥2 as the cut-off. For strand specific data, we obtained overall low reads from both biological replicates compared with double strand sequencing. Hence a lower cut-off was used (Table 1). Since the read count for silent or low-expressing genes was zero, the cut-off value of 2 was considered appropriate for expressed genes. One copy had read count <2.0 in both sense and antisense directions. Of the remaining 40 copies, 28 showed antisense transcription while 37 copies showed sense transcription. Three copies were exclusively transcribed in antisense direction while 12 copies were exclusively transcribed in sense direction (Table 1). These data showed that sense transcription was more common than antisense in terms of number of transcribed copies, although the overall number of antisense reads was about two-fold higher than sense reads (Fig. 6). The high antisense read count was mainly due to two

Table 1
Expressed EhLINE1 copies of *Entamoeba histolytica*.

S.No.	Scaffold	Position in Scaffold	Length of LINE copy (bp)	FL / Truncation (#)	log2_average -pme_TPM	Sense-average-expected_count		Antisense-average -expected_count
						EhLINE1	RT alone	
1	[b]*DS571495	5888..10696	4809	Full Length	6.97	13.455	6	295.72
2	*DS571290	32725..37508:r	4784	Full Length	5.31	4.620	0	98.86
3	DS571155	84542..84887:r	346	5' truncated (4438-end)	5.11	5.500	NA	0.00
4	DS571261	37357..37932	576	5' truncated (4207-end)	4.90	5.000	NA	0.00
5	DS571201	32751..33214	464	5' truncated (4310-end)	4.60	0.000	NA	4.00
6	DS571267	21652..22101	450	Both end truncation (2630 - 3076)	4.50	5.500	5.5	0.00
7	DS572240	698..1493:r	796	5' truncated (3986-end); at scaffold end	4.32	2.490	0	20.39
8	DS571337	29259..30063	805	3' truncated (start-1107)	3.99	41.500	NA	0.00
9	DS571493	867..2479:r	1613	5' truncated (3178-end)	3.95	3.500	0	11.56
10	[a]*DS571192	10681..15465	4785	Full Length	3.88	10.110	2	43.90
11	DS571181	1..2060	2060	5' truncated (2719 - end); at scaffold start	3.51	5.645	3	7.82
12	[EN]DS571396	357..2911:r	2555	5' truncated (2237-end) RT	3.46	6.500	3	12.77
13	[R]DS571271	2766..6879	4114	D_EhL1	3.27	7.000	2	25.01
14	*DS571606	4111..8079:r	3969	3' truncated (start-3976)	3.20	6.975	1	18.32
15	DS571570	2575..5614:r	3040	Both end truncation (1375 - 4410)	3.03	3.575	3	12.91
16	[R]DS571274	20255..22076:r	1822	D_EhL1	2.92	13.500	NA	0.00
17	[R$]DS571214	55871..57424	1554	D_both_end_truncation (879-4460)	2.91	19.235	NA	0.50
18	[R]DS571236	41813..44330	2518	D_EhL1	2.91	25.500	NA	0.50
19	[a]*DS571151	66192..70991	4800	Full Length	2.91	10.420	3	23.94
20	[R]DS571432	2548..3739	1192	D_3' truncated (1-1505)	2.80	12.000	NA	0.50
21	[R]DS571219	29209..33989	4781	Full Length	2.62	13.465	2	14.68
22	[$]DS571387	5598..10374	4777	Full Length	2.58	9.160	0	21.50
23	*[RT]DS571434	2047..6842	4796	Full Length	2.58	3.415	2	22.50
24	*DS571467	4357..9130:r	4774	Full Length	2.55	2.970	0	14.31
25	DS571269	38394..43132:r	4739	Full Length	2.53	8.000	0	11.50
26	DS571158	57055..61833:r	4779	Full Length	2.44	20.080	4	9.99
27	[R]DS571373	10757..12687:r	1931	D_5' truncated (499 - 4781)	2.41	19.500	NA	0.56
28	[R]DS571351	19528..21607:r	2080	D_EHL1	2.36	11.000	NA	0.00
29	DS571193	25627..30408:r	4782	Full Length	2.29	10.200	3	11.80
30	[R]DS571312	21378..24390:r	3013	D_5' truncated (230 - end)	2.23	3.345	0	8.05
31	*DS571417	7568..12350	4783	Full Length	2.12	6.640	1	10.29
32	DS571282	33548..38335:r	4788	Full Length	1.85	12.510	6	8.00
33	DS571362	9194..13984	4791	Full Length	1.69	14.015	0	6.08
34	DS571218	1012..5800	4789	Full Length	1.61	4.985	0	3.97
35	[R]DS571477	8414..12103:r	3690	D_EhL1	1.42	1.000	0	1.00
36	DS571529	5122..9932	4811	Full Length	1.23	1.020	0	2.93
37	[R]DS571186	60813..63836:r	3024	D_5' truncated (222-end)	1.18	6.955	0	1.00
38	[R]DS571304	4639..9366	4728	Full Length	0.81	1.765	0	3.02
39	[a]DS571160	55138..59914	4777	Full Length	0.80	6.335	0	6.00
40	[R]DS571372	19954..24632:r	4679	Full Length	0.77	3.415	0	4.50
41	DS571233	48784..53550:r	4767	Full Length	0.67	4.465	0	1.01

Symbols Used
a : EhLINE1 with intact ORF1 reading frame
b: EhLINE1 with intact ORF2 (has both EN and RT reading frame intact)
EN: Intact EN reading frame
RT: Intact RT reading frame
$: Readthrough transcription
R : Re-annonation
* : Validated by RT-PCR
:r : Reverse complement
D_EhL1: EhLINE1 with both ends intact but internal deletions
D_5' truncated: EhLINE1 with 5' end truncation and has internal deletions
D_3' truncated: EhLINE1 with 3' end truncation and has internal deletions
D_both_end_truncation: EhLINE1 with both ends trauncation and also has internal deletions
NA: Not applicable
#: Nucleotide Position w.r.t full length EhLINE copy of scaffold DS571192

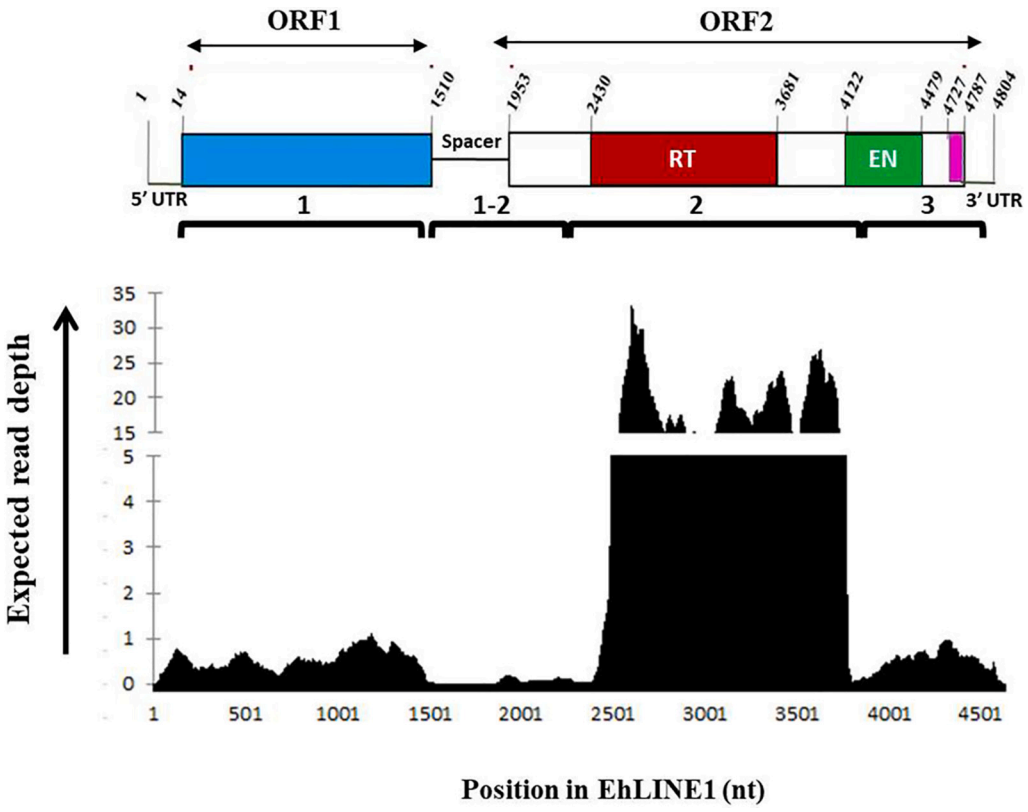

Fig. 2. RNA-Seq reads from different regions of EhLINE1. The domains of EhLINE1 are marked according to positions in the full-length copy at locus DS571495. ORF1 extends from nucleotide position 14 to 1510 and ORF2 from position 1953 to 4784. Within ORF2 the location of RT and EN domain is marked. A 74 bp stretch at 3′-end of EhLINE1 (shaded in pink) shares sequence similarity with 3′-end of EhSINE1. RNA-Seq reads coming from this region were not considered. The "spacer" is the sequence between stop codon of ORF1 and first AUG of ORF2. Wiggle plot shows the pattern of RNA-Seq reads along the length of EhLINE1. 'Expected read' denotes the maximum likelihood abundance estimate. Data shown is average of all 20 expressed full-length EhLINE1 copies from three replicates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

copies which showed very high read count (Table 1) compared with transcription in the sense direction from any copy. Since most of the antisense transcription came from the RT region, we checked the status of sense transcription of this region, especially from full-length copies that contained the entire sequence. Of the 20 full-length copies, 19 showed antisense transcripts while only 8 copies showed sense transcripts from the RT region. The copy with the highest sense RT expression also showed the highest antisense expression. Only one copy had no antisense expression, but showed sense RT expression. However, this copy (DS571267) was truncated at both ends, with a size of only 446 bp that included part of the RT region. Thus, RNA-Seq data revealed for the first time the presence of antisense transcripts primarily from RT region of EhLINE1.

The strand-specific expression data were experimentally validated by northern analysis using single-stranded probes from ORF1 and RT regions. The sense-strand probes from both regions hybridized with 1.5 kb bands, as expected. No signal was observed with antisense probe for ORF 1. However, strong signal of 1.5 kb RNA was seen using antisense RT probe (Fig. 7). This further confirmed the presence of antisense RT transcripts. Another line of evidence was obtained for expression of antisense RNA by checking the EST database. Two ESTs "CX099071.1 and CX095831.1" mapped with 100% identity in antisense orientation to the full-length EhLINE1 copy (DS571495, Table 1), which showed maximum antisense expression in our RNA-Seq data. The sequence match was for nt position 2432 to 3155, which lies within RT domain (Additional File 2: Fig. S8). The expected 3′-end of the antisense RNA reads in RT region (at nt position 2440) matched with the 3′-ends of the ESTs CX099071.1 and CX095831.1 at positions 2432 and 2433 respectively. These data, taken together, strongly suggest the presence of a novel long antisense RNA corresponding to the RT region of EhLINE1.

To explore the translational potential of this antisense RNA we looked for possible ORFs using expasy translate tool (Gasteiger et al., 2003) and found 4 non overlapping peptides of sizes 150, 47, 42 and 28 aa (Additional File 2: Fig. S9). We checked codon usage of the peptides through the codon usage program of Sequence Manipulation Suite (Stothard, 2000) to determine frequency of low-usage codons (Additional File 3: Table S1, Additional File 4). Compared with sense sequences of EhLINE1 ORF1p and ORF2p, the antisense peptides showed significantly greater frequency of low-usage codons. We searched all available protein databases (including AmoebaDB) for matches with the peptides coming from antisense transcript, using Quick BLASTP, blastp, psi-blast and delta-blast. No significant hits were obtained. Thus, we could not obtain evidence for possible translation of the RT antisense RNA into functional peptides/proteins. However, due to low coverage of available *E. histolytica* proteomic data we were also not able to find peptides corresponding to EhLINE1 ORF1p, which we could otherwise detect by western blotting (Yadav et al., 2012). This question needs to be resolved with more sensitive analysis.

3.4. Translation potential of transcribed EhLINE1 copies

Having obtained sense strand-specific transcription data for individual copies we checked which copies could potentially get translated by looking for intact reading frames corresponding to ORF1 and ORF2. Of 742 EhLINE1 copies no copy was found with both the ORFs intact. Six copies had intact ORF1 while only 1 copy had intact ORF2 (Table 1). Of the 6 copies with intact ORF1, only 3 were transcriptionally active, as no reads mapped to the other three copies. All three transcribed copies were full-length (locus DS571192, DS571151 and DS571160), and all showed comparable level of sense strand expression for ORF1 but no antisense ORF1 transcripts (Additional File 2: Fig. S10). The single copy with intact ORF2 reading frame (locus DS571495) was also full-length. Interestingly, it contributed to the maximum number of antisense transcripts from RT region (Table 1). The sense strand of this copy showed transcripts from all the three regions. Further we looked for copies (apart from this ORF2 copy) with intact reading frames for either RT or EN domains individually. We found one full-length EhLINE1 copy (locus DS571434) which had an intact RT reading frame. It showed RT

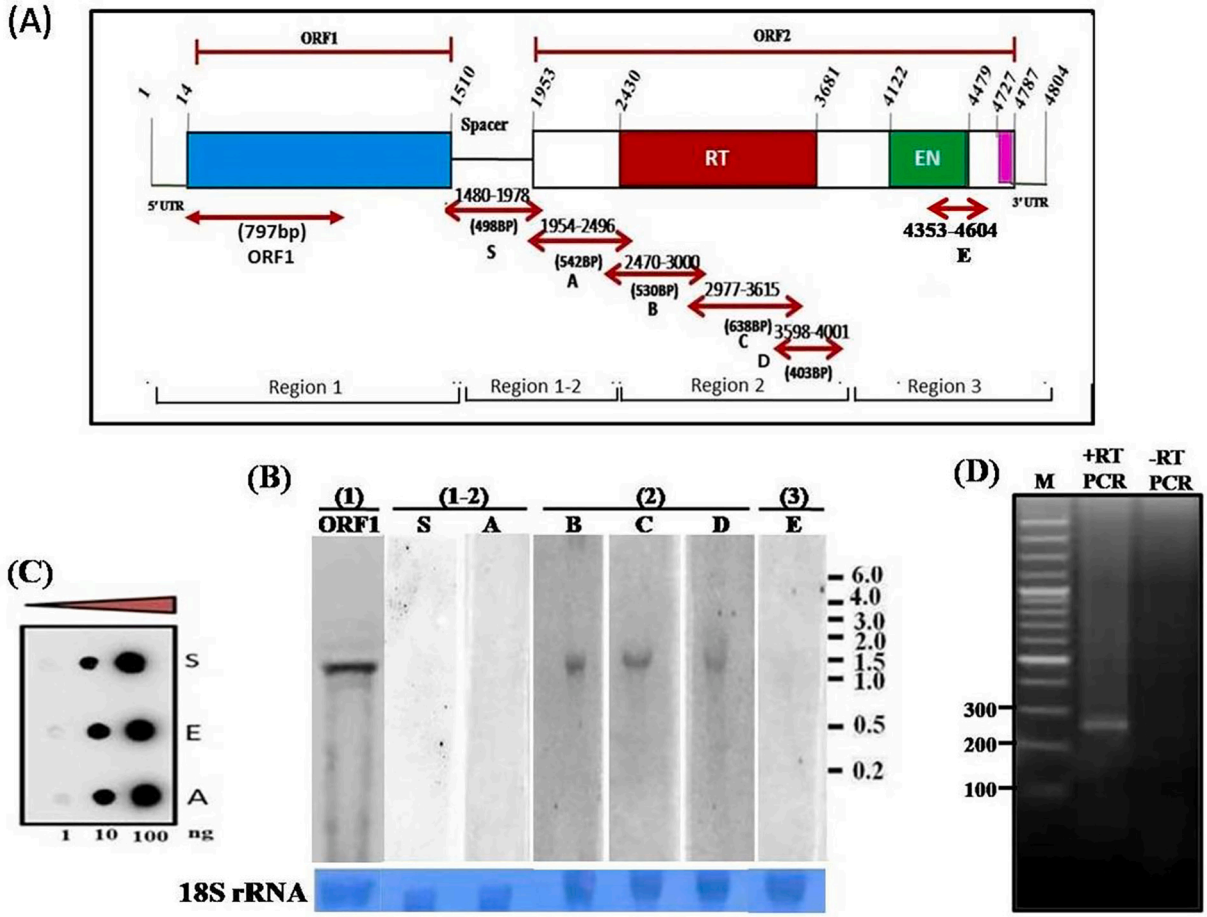

Fig. 3. EhLINE1 expression measured by northern blot and RT-PCR. (A)Probes (ORF1, S, A, B, C, D, E) were designed from different regions of EhLINE1. (B) Northern analysis with the above probes. 18S rRNA (methylene blue stained) was used as loading control. No signal was detectable from region (1–2) and region 3. (C) The quality of probes 'S', 'A' and 'E' which did not give any signal in northern blot, was confirmed by dot blot analysis with increasing concentration of DNA. (D) Transcripts from region 3 were detectable by RT-PCR. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reads in both sense and antisense direction, although expression in the sense direction was low. Intact EN reading frame was found in 6 EhLINE1 copies (apart from DS571495). Of these, RNA-Seq reads mapped only to 1 copy (locus DS571396) (Table 1). It was a 5′-truncated copy spanning nt position 2237 till 3′-end. The expression from EN region was low in both copies, and reads were from sense strand alone.

To sum up, our analysis showed that potentially translatable ORF1 was transcribed from three copies exclusively in sense direction. No potentially translatable copy with ORF2 or RT domain was transcribed in sense direction alone. There was extensive antisense transcription of the RT domain. One copy with EN domain was transcribed in sense direction alone at a low level. These factors explain our earlier observation that *E. histolytica* cells express ORF1p, but ORF2p was undetectable (Yadav et al., 2012). The role, if any, of antisense RNA in inhibiting the translation of RT domain remains to be studied.

3.5. Absence of full-length EhLINE1 sense transcripts

RNA-Seq data showed that 20/41 transcribed EhLINE1 copies were full-length. Yet we failed to see any full-length (4.8 kb) transcripts both in northern blots (Fig. 3), and from RNA-Seq data. We therefore looked for internal promoters and polyadenylation sites. It is generally believed that LINE elements are transcribed into polycistronic mRNAs from a promoter located at the 5′-end (Heras et al., 2007; Macías et al., 2016). We checked to see if such a promoter was active in EhLINE1 and whether a second promoter existed upstream of RT domain that could be responsible for the 1.5 kb transcript from this region. Promoter activity was measured by a luciferase reporter assay using fragments cloned upstream of luciferase. A 200 bp DNA fragment from 5′-end of EhLINE1 showed promoter activity, while no activity was seen in a fragment that included only 100 bp from 5′-end (Fig. 8). This showed that, similar to LINE elements in other organisms, EhLINE1 had an internal promoter within 200 bp at 5′-end. To check whether a second promoter existed upstream of RT domain we cloned a 1489 bp fragment (1511 nt-3000 nt) upstream of luciferase. However, no luciferase activity was obtained with this fragment (Fig. 8). Further work is required to understand how the 1.5 kb RT domain transcript (and the EN domain transcript) are generated.

Many of the truncated transcripts of human L1 element have been shown to correspond to internal polyadenylation sites in the endogenous elements (Perepelitsa-Belancio and Deininger, 2003). We checked for the presence of such sites in EhLINE1. The consensus polyadenylation signals of *E. histolytica* genes have been well documented from a number of studies (López-Camarillo et al., 2005; Zamorano et al., 2008). Three sequence elements reported in this context include the consensus AAWUDA motif (polyadenylation signal) located 20 nt upstream of polyadenylation site, enrichment of C at −1 nt position, and a broad T-rich region surrounding the polyadenylation site (Hon et al., 2013). We checked to see if such motifs indicative of polyadenylation existed throughout the EhLINE1 sequence. We found 23 sites with consensus polyadenylation signal motif but only 2 sites met all the three criteria (Additional File 5). These 2 sites corresponded to nt positions 1510 (end

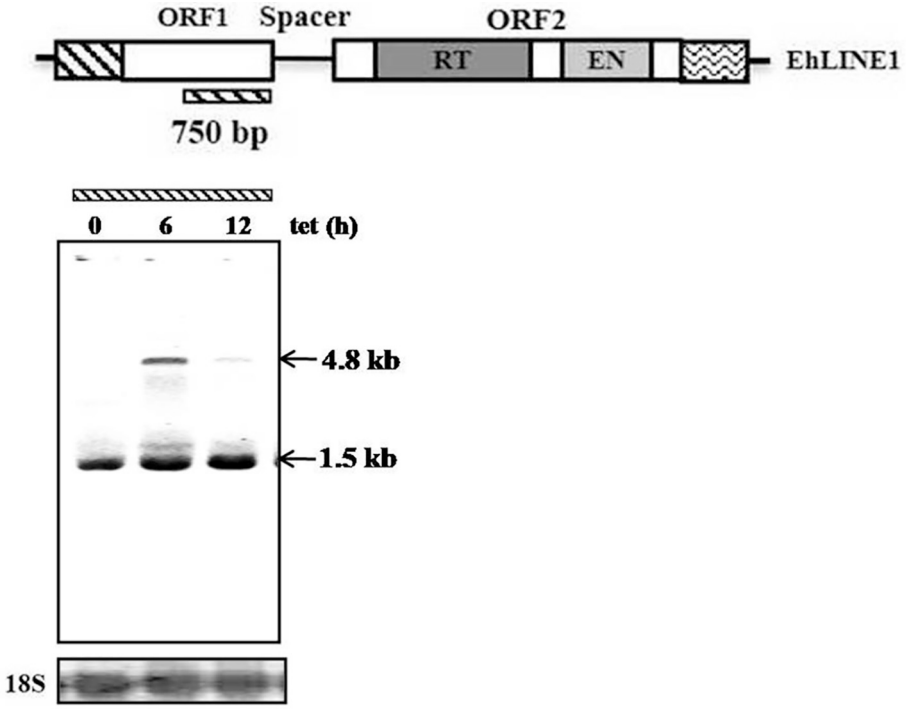

Fig. 4. Northern hybridization with total RNA from *E. histolytica* cells transfected with 4.8 kb EhLINE1 copy expressed from a tetracycline-inducible promoter. Cells were induced with 10 μg/ml tet for the indicated times. DNA probe of ORF1 (750 bp) was used. 18S rRNA, used as loading control, was stained with methylene blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

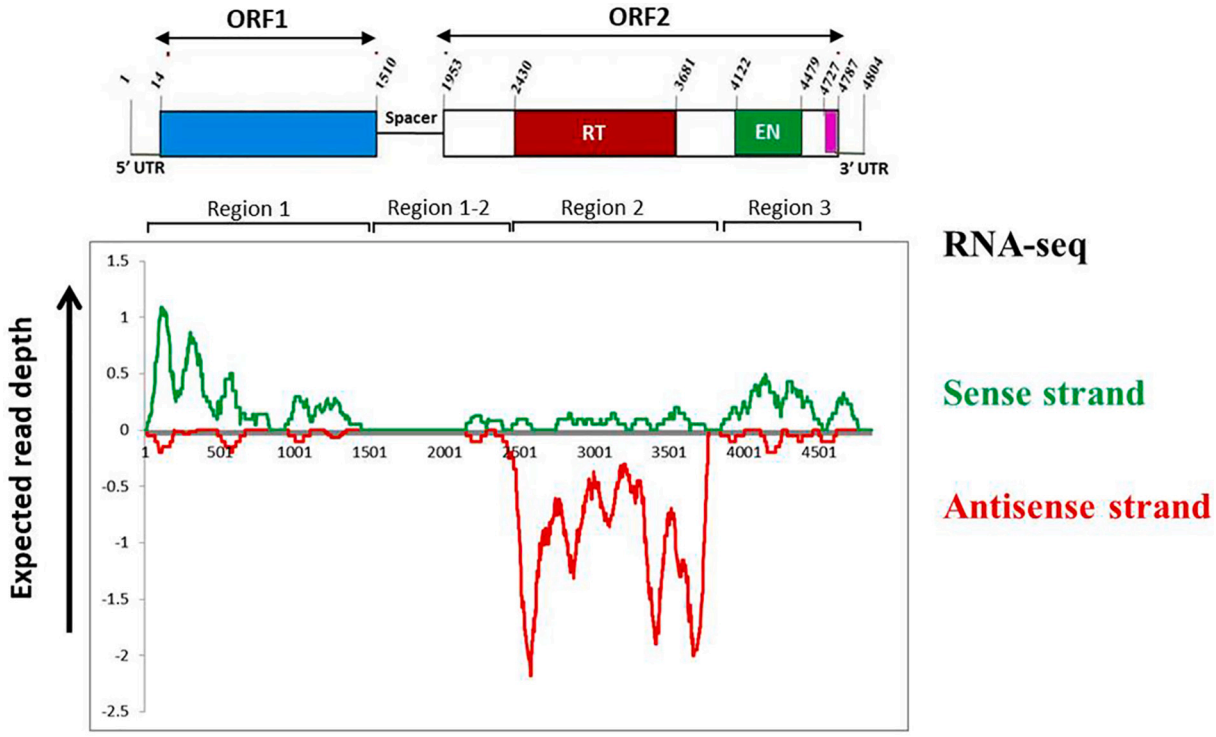

Fig. 5. Distribution of sense and antisense reads along the length of EhLINE1. Graph shows the average of all 20 expressed full-length EhLINE1 copies. Antisense transcripts were predominantly seen from RT region. Very few antisense reads mapped to ORF1 and region 3. 'Expected read' denotes the maximum likelihood abundance estimate.

of ORF1), and 4770 (end of ORF2) of EhLINE1 copy in DS571192. They were also found in the other expressed full-length copies at the corresponding positions (Additional File 5), and in copies with intact ORF1 (loci DS571160, DS571151 and DS571192; Table 1) or intact ORF2 (locus DS571495) (Fig. 9). We also examined the sequence at the end of RT domain in all full-length expressed EhLINE1 copies and could locate the consensus polyadenylation signals, although the T-rich stretch was poorly defined (Additional File 5: Table S2, Fig. S17). Further, we

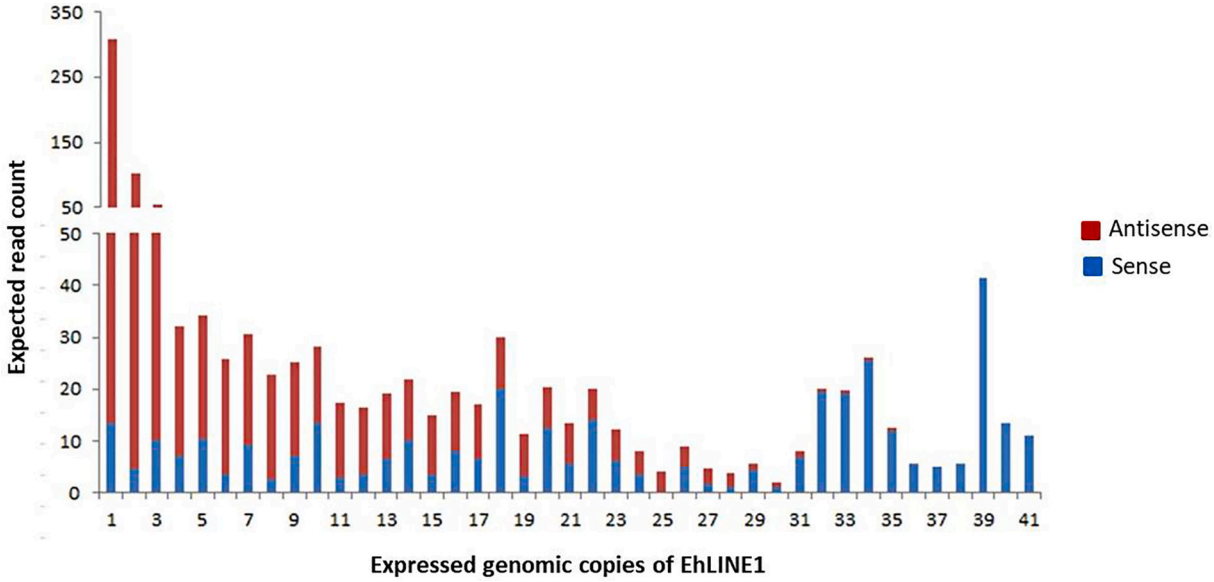

Fig. 6. Ratio of sense vs. antisense reads in EhLINE1 expressed copies. The data was plotted from Table 1. The EhLINE1 copies are shown in decreasing order of expression from antisense strand.

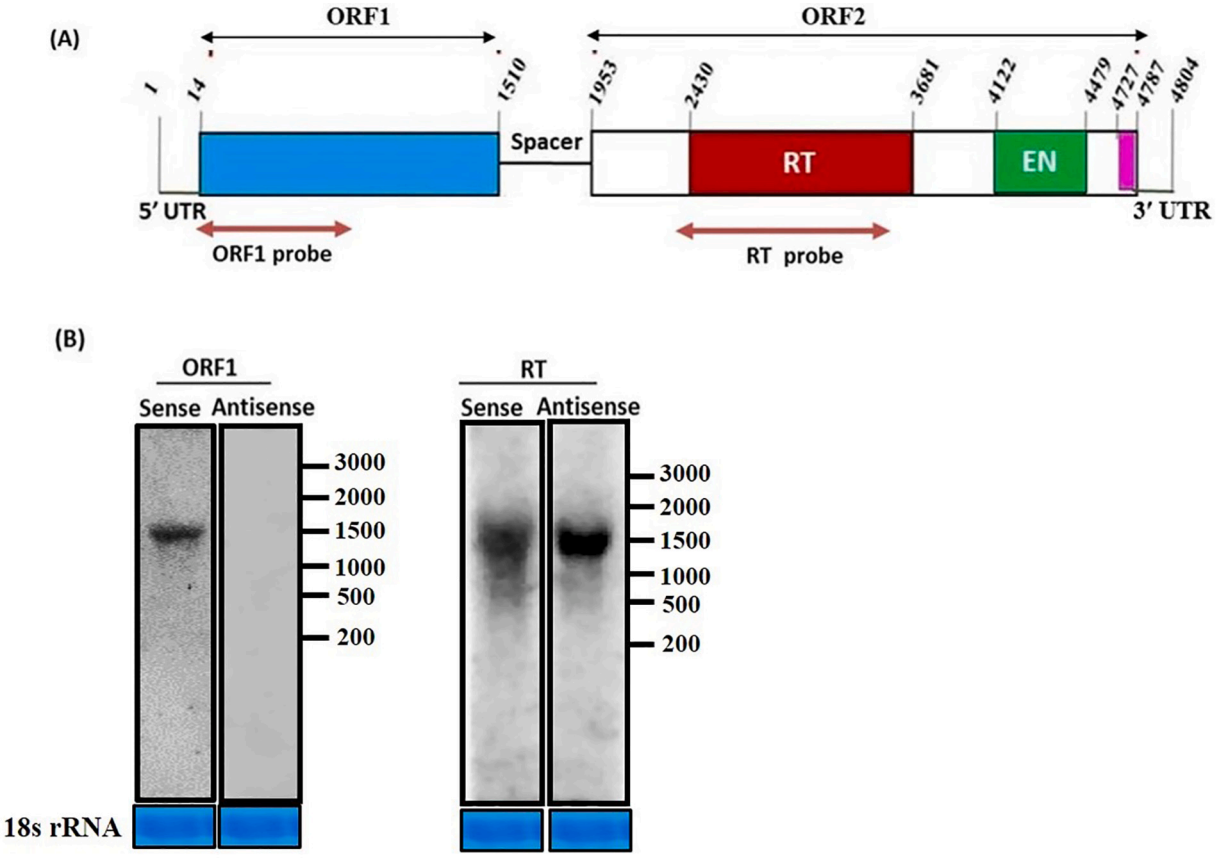

Fig. 7. Northern analysis of antisense transcription of EhLINE1. (A) Location of sense and antisense probes from ORF1 and RT are shown in EhLINE1. (B) Northern analysis with sense and antisense probes. Size markers (bases) are shown on the left. 18S rRNA stained with methylene blue served as loading control. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

experimentally confirmed that the ORF1 and RT transcripts were polyadenylated by RT-PCR analysis of total RNA using oligo(dT) primer for reverse transcription followed by PCR with ORF1- and RT-specific primers (Fig. 10).

These data suggest that ORF1 is transcribed from the promoter located at 5′-end of EhLINE1, and the transcript is possibly polyadenylated using consensus sequence motifs at the ORF1 3′-end. The 1.5 kb RT transcript could also be polyadenylated using consensus motifs. However, it is not clear whether it is transcribed from an internal promoter (which we could not detect), or is processed from a polycistronic

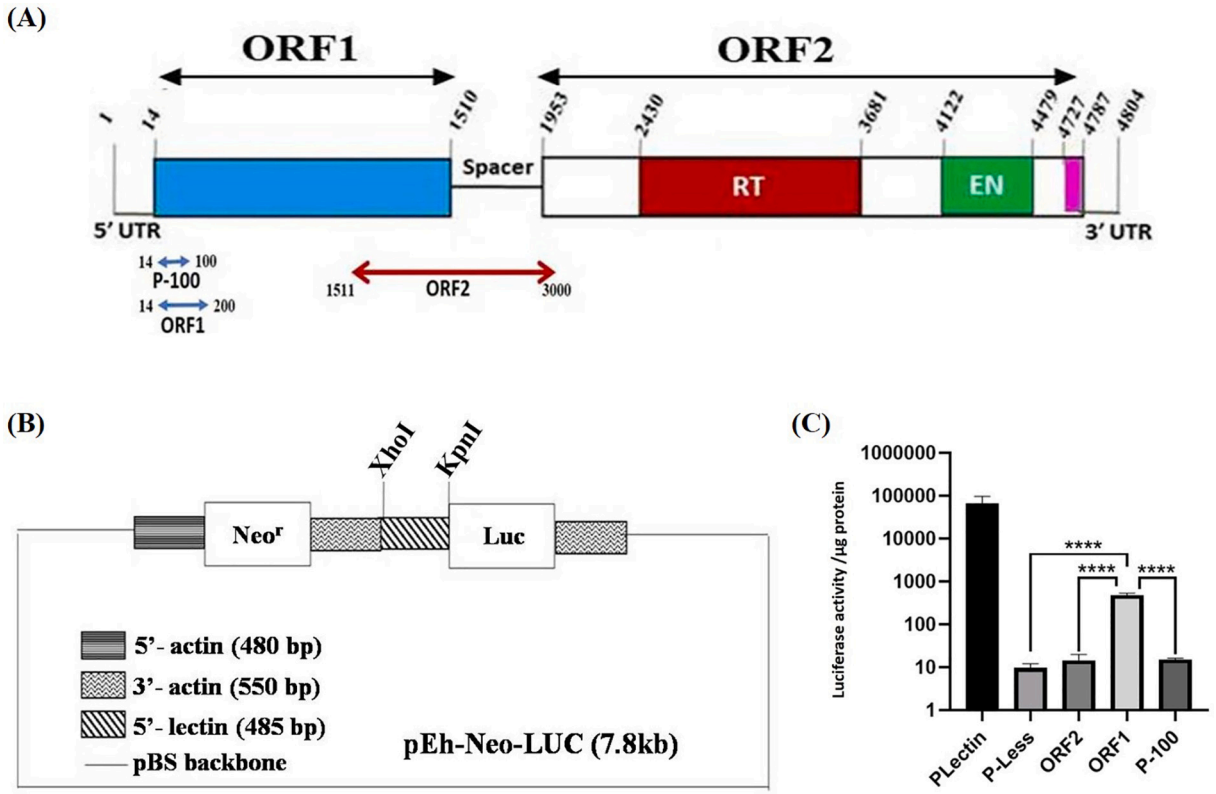

Fig. 8. EhLINE1 promoter analysis by luciferase (LUC) expression assay. (A) The fragments tested for promoter activity are indicated in EhLINE1. P-100 (14–100 bp); ORF1 (14–200 bp); and ORF2 (1511–3000 bp). (B) Map of the pEh-Neo-LUC vector used for cloning the above fragments upstream of LUC. (C) Luciferase reporter assay with lysates from stable transfectants. P-Less, which lacked any promoter fragment, served as negative control, whereas P-Lectin (with *E. histolytica* lectin promoter) was a positive control. The data are average of three independent measurements. The difference in variance was checked using ANNOVA which showed *P*-value = 0.0003. The mean values (±S.D.) are shown. Difference in means was represented as an asterisk, where four asterisks represent p value < 0.0001.

transcript.

3.6. Contribution of flanking sequences to the transcription of EhLINE1 copies

In addition to transcription originating from the EhLINE1 promoter at 5′-end it is possible that some LINE copies may get transcribed by read-through transcription from neighbouring genes. To determine this, in both sense and antisense directions, for all the 41 expressed EhLINE1 copies, we took 200 nt sequence from the end of each copy and 200 nt from the end of nearest flanking gene, together with the intergenic region (the intergenic regions in *E. histolytica* are generally very short; (Bruchhaus et al., 1993; Petter et al., 1992; Willhoeft et al., 1999) and looked at the RNA reads mapping in this region. Only 2/41 copies showed some read through transcription, at a very low level (Additional File 2: Fig. S11, S12). One of these was a full-length copy (locus DS571387), with midasin gene located 156 bp from its 3′-end. The other copy was "both ends truncated", (locus DS571214) and the gene Sec61 alpha subunit (putative), was located 67 bp upstream of it. The data showed that the contribution of read-through transcription to the EhLINE1 transcriptome was not significant.

Further, we checked for possible contribution by promoter elements located upstream of EhLINEs. We looked at 13 expressed copies that lacked the 5′-end (5′ truncated and both ends truncated). To look for possible upstream *E. histolytica* promoter elements in these copies, we extracted 200 bp sequence 5′-upstream of each copy of EhLINE1 and searched for *E. histolytica* specific promoter motifs (Purdy et al., 1996), using FIMO (Grant et al., 2011) and MAST (Bailey and Gribskov, 1998). For two copies (DS571186 and DS571493) TATA box motifs were found at position −125 and −85 respectively, upstream of EhLINE1 5′-end (Additional File 2: Fig. S13A). However, no RNA-Seq reads originated from these positions.

In two other copies (DS571267 and DS571201) TATA box and Inr box could be located within 250 bp upstream in the sense or antisense orientation, respectively (Additional File 2: Fig. S13B, S14; S13C, S15). However, DS571201 lacks the RT region and contributes to very few antisense reads (Table 1).

Thus, our data showed that read-through transcription, or expression from upstream promoter elements might have a very limited role in driving the expression of EhLINE1. Another possibility could be that EhLINE1 transcripts might originate due to splicing from transcripts of neighbouring genes. Although introns are not common in *E. histolytica* mRNAs, we tested this possibility by mapping all the splice junctions from RNA-Seq data using HISAT2 (Kim et al., 2019). Only one splice junction mapped to EhLINE1 copy (DS571181). Thus, we could not find evidence of spliced reads from other genes into EhLINE1 copies.

4. Discussion

The steady state levels and sizes of LINE transcripts from endogenous genomic copies in different organisms is highly complex and indicates that LINE transcription is subject to a number of regulatory factors. These likely operate at different levels, including selection of transcriptionally active copies, use of alternative promoters, polyadenylation sites and other processing events, and read-through transcription. Due to the very large copy number of LINEs in most organisms it has been difficult to obtain a comprehensive picture of the transcription status of all genomic copies. We have attempted such a study in *E. histolytica* where the EhLINE1 copy number is relatively small. Our RNA-Seq analysis showed that the pattern of EhLINE1

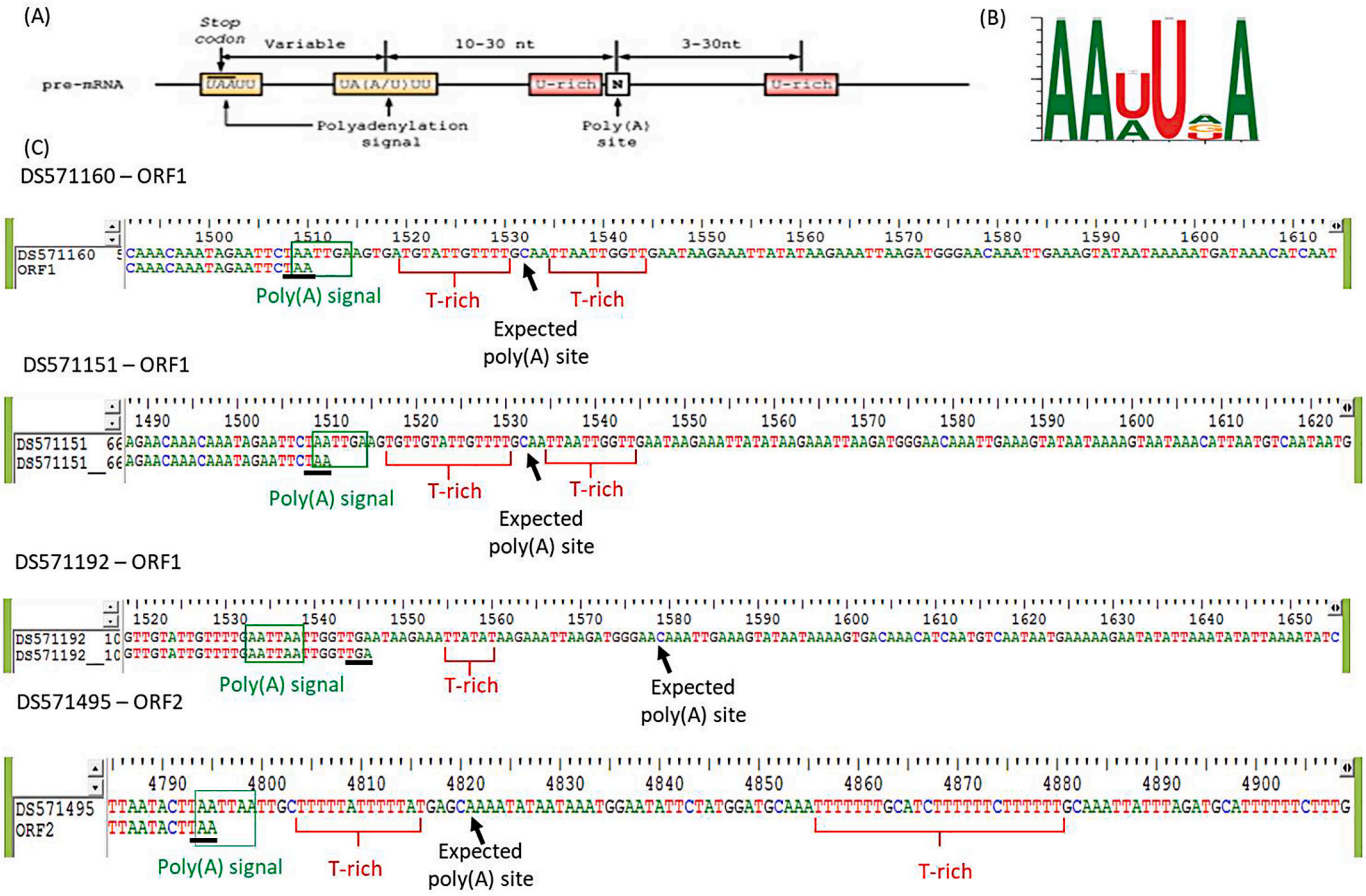

Fig. 9. Consensus polyadenylation signals at 3′ –ends of EhLINE1 ORFs. (A) Characteristic features of polyadenylation signal in *E. histolytica*, adapted from (Zamorano et al., 2008). (B) AAWUDA motif (polyadenylation signal) in *E. histolytica*, adapted from (Hon et al., 2013). (C) Polyadenylation consensus features at 3′-end of ORF1 or ORF2 in EhLINE1 copies with each complete ORF. Poly(A) signal- boxed in green, T-rich region- highlighted in red, stop codons- black bar and expected poly(A) site-solid arrow are indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

transcription shares several common features with other LINEs. Firstly, EhLINE1 contains an internal promoter located within 200 bp of the 5′-end, which can support sense strand transcription. Such a promoter, driven by RNA polymerase II is reported in many LINEs, including human L1(Alexandrova et al., 2012; Minakami et al., 1992; Swergold, 1990), and *Drosophila* (Mizrokhi et al., 1988). Among protistan parasites, the *Trypanosoma cruzi* LINE, L1Tc also has an internal promoter (Pr77) located within 77 bp at the 5′-end (Heras et al., 2007). It is highly conserved in all retrotransposons of *T. cruzi*, and has important sequence motifs required for transcription (Macías et al., 2016). Secondly, only a small fraction of EhLINE1 copies (6.7%) are transcriptionally active. About half of the transcribed copies are full-length. Most human L1 copies are also found to be transcriptionally inactive. A detailed analysis of expressed sequence tags corresponding to human L1 in a lymphoblastoid cell line showed evidence of transcription at only 692 L1 element sites, of which 410 were full-length (Rangwala et al., 2009). The retrotranspositionally most active human L1 members belong to the L1 HS-Ta subfamily. It was shown that the L1HsTa copies which were highly expressed were located in loci that were also highly expressed, and the environment in which the element was inserted had a strong influence on expression of LINE copy (Philippe et al., 2016). We analyzed *E. histolytica* RNA-Seq data (Naiyer et al., 2019) to look at expression levels of genes flanking the 41 EhLINE1 expressed copies. We could retrieve this information for 28 copies (the rest being at the end of scaffolds), of which we found only 3 that were flanked by genes in the higher expression categories, while the remaining were medium or low-expressing (Additional File 6: Table S3). The maximally expressed EhLINE1 copies were not flanked by highly expressing genes. Thus, we did not see any correlation between expression levels of flanking genes with that of the neighbouring EhLINE1 copy. Thirdly, most EhLINE1 transcripts in our study (from exponentially growing cells) were truncated. We tested RNA from cells grown in a variety of conditions, e.g. subjected to growth stresses like heat shock and oxygen stress. In northern blots these RNAs also failed to show full-length EhLINE1 transcripts (Additional File 2: Fig. S16). However, it is possible that under certain conditions EhLINE1 may be retrotranspositionally activated and transcribed frequently into its full-length RNA. In both human and mouse cells most LINE transcripts are truncated, with full-length transcripts seen only in some cell types (Dudley, 1987; Martin, 1991; Packer et al., 1993). In human L1 the truncation sites have been shown to correspond with internal polyadenylation sites, and this could be a mechanism to limit the production of full-length retrotranspositionally-competent transcripts (Perepelitsa-Belancio and Deininger, 2003). Short transcripts have also been reported in the major LINE (L1Tc) of *T. cruzi*. Northern analysis of L1Tc showed a major band of 5 kb along with shorter transcripts, especially 0.4 kb and 0.2 kb (Trelogan and Martin, 1995).

Apart from patterns generally shared with other LINEs, EhLINE1 transcripts displayed some interesting unique features. The sense strand transcripts, observed both by RNA-Seq and northern hybridization, corresponded to the major functional domains of EhLINE1 (ORF1, RT and EN). This was not related to truncations of EhLINE1 copies at the DNA level since DNA truncations showed no correspondence with functional domains, and 49% of the transcribed copies were, in fact, full-

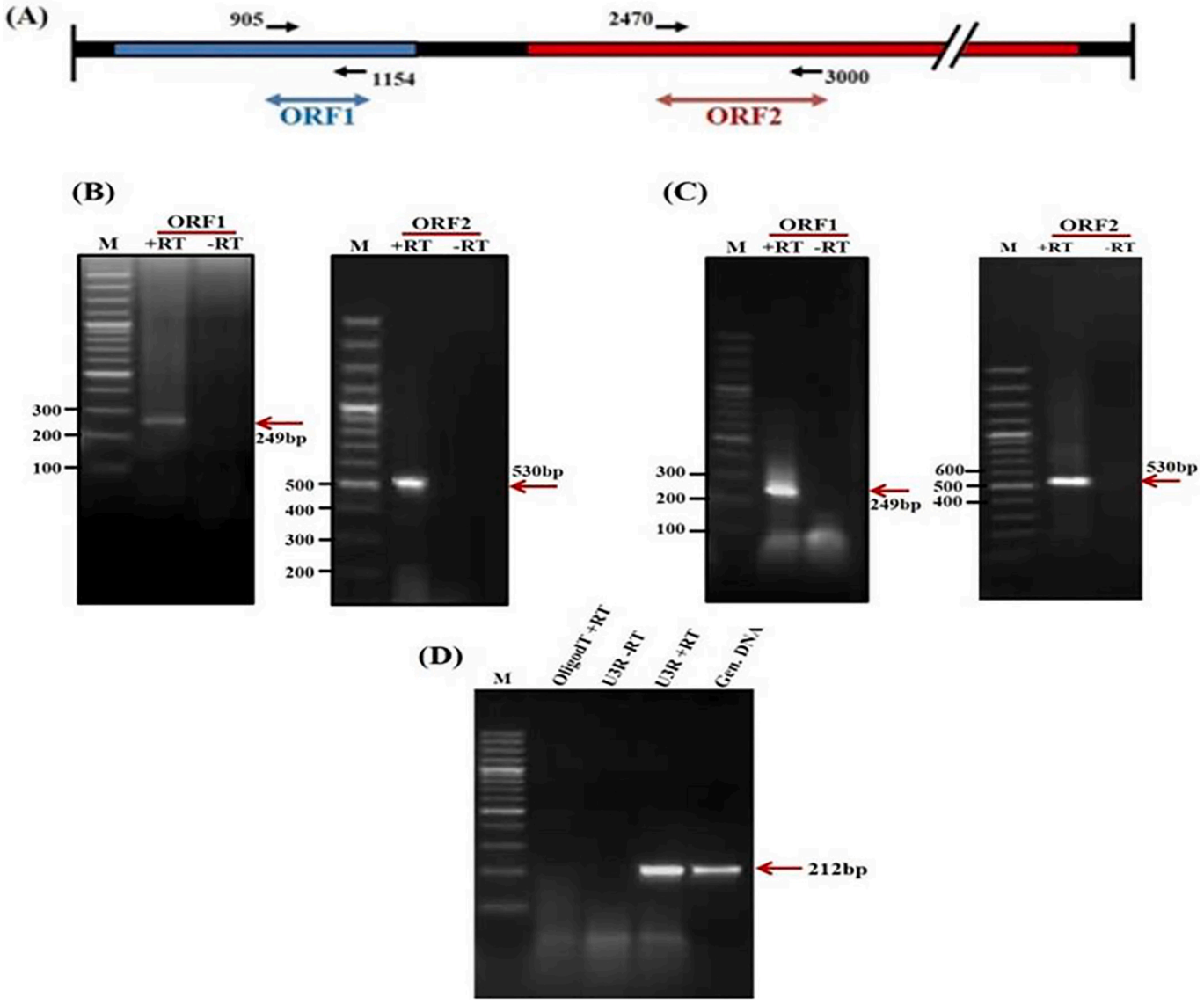

Fig. 10. Polyadenylation status of ORF1 and ORF2 transcripts. cDNA was synthesized using a 45-mer oligo(dT) primer at 50° C (since *E. histolytica* genome is A + T rich). (A) Location of primer pairs used for PCR. (B) DNase treated total RNA (5 μg) was used for cDNA synthesis with oligo(dT) primer, followed by PCR with ORF1 and RT specific primers. Amplicons were checked by 1.5% agarose gel electrophoresis. (C) As in (B), using poly(A) + RNA (500 ng) for cDNA synthesis. (D) U3 snoRNA which is not polyadenylated was used as negative control. U3 primers gave no amplicon with cDNA made with oligo(dT) primer while amplicon was obtained with U3 specific reverse primer (U3R).

length (including the top two highly transcribed copies; Table 1). The genesis of truncated EhLINE1 transcripts is not clear at present. Whether these transcripts are obtained by internal processing from a single full-length transcript initiating from the promoter at 5′-end, or are generated by transcription initiation from alternative promoters remains to be resolved. Internal processing of precursor polycistronic transcripts by trans-splicing of spliced leader sequences is a well-known mechanism to generate functional mRNAs in a variety of organisms, including parasitic protists like *Trypanosoma* and *Leishmania* (Hastings, 2005; Vesteg et al., 2019). Although trans-splicing has not yet been reported in *E. histolytica*, it is possible that such a mechanism could generate the truncated EhLINE1 transcripts observed by us. Other alternatives like splicing from read-through transcripts appear unlikely. Whatever might be the mechanism, it is significant that transcripts of the functional domains are present in these cells. The activities required for retrotransposition could be available if these transcripts from ORF1, RT and EN domains are translatable. The ORF1 polypeptide is constitutively expressed in *E. histolytica* (Yadav et al., 2012), but ORF2 could not be detected. If the RT and EN domains are translated under certain conditions, it is possible that these activities could mobilize EhLINEs, or EhSINEs *in trans*. We have earlier shown that EhSINE1, the nonautonomous partner of EhLINE1 with which it shares the 3′-end (Bakre et al., 2005), could be mobilized by ectopic overexpression of EhLINE1 ORF2 (Yadav et al., 2012). A role for trans-mobilization of truncated LINE transcripts has been suggested in mouse (Branciforte and Martin, 1994).

The most important unique feature we found in EhLINE1 was the strong antisense transcription specifically from RT domain. The presence of antisense RT transcripts was clearly demonstrated both by mapping of RNA-Seq reads and by northern hybridization, and was also corroborated by matches with EST database. At present we are unable to comment on the mechanism by which these transcripts are generated. This will be taken up in future studies. However, the existence of such transcripts is a significant novel observation. Although antisense transcription is very well defined in human L1 it is fundamentally different from that observed by us. The L1 antisense promoter (ASP) is located in the 5′-UTR between nucleotides 400 to 600 (Speek, 2001). The primate-

specific LINE1 contains an open reading frame termed ORF0 between nucleotides 452–236 in the antisense orientation. It lies downstream from the ASP, has a strong, well-conserved Kozak sequence and is translated (Denli et al., 2015). In addition, transcription from ASP in L1s inserted in introns of genes in antisense orientation can result in the formation of chimeric transcripts (Criscione et al., 2016). By contrast, antisense transcripts in EhLINE1 are predominantly derived from the region that, in sense orientation, encodes the RT domain. They are large transcripts, being 1.5 kb in size (as confirmed in northern blots), with distinct ends mapped by RNA-Seq reads. From sequence analysis they do not seem to encode any known functional peptides. Another striking feature is that the overall number of antisense sequence reads exceeds sense reads in EhLINE1. Such massive antisense transcription has not been previously reported in LINEs. The EhLINE1 promoter/s from which these antisense transcripts originate have yet to be mapped. They could be located within ORF2. Interestingly, the mammalian LINE-1 ORF2 has been shown to promote both sense and antisense transcription during neuronal differentiation. The ORF2 region was found to contain a number of overlapping Sox/LEF binding sites. These could promote transcription in both directions mediated by Wnt/β-catenin activation (Kuwabara et al., 2009).

We believe that antisense transcripts in EhLINE1 could be involved in attenuating translation of the RT domain, as we failed to detect ORF2 polypeptide in *E. histolytica* cells. On the other hand, ORF1, for which we did not find any antisense transcripts, was translated at high levels (Yadav et al., 2012). An effect of antisense RNAs on reducing the level of proteins required for replication and integration of LTR retrotransposon Ty1 has been suggested in *S. cerevisiae*. Antisense RNAs of size 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. PIWI-interacting small RNAs are known to repress transposons by transcriptional or posttranscriptional mechanisms in a variety of organisms (Iwasaki et al., 2015). The RNAi pathway has been well characterized in *E. histolytica*. The predominant sRNAs in *E. histolytica* are 27 nt with a 5′-polyP structure (Zhang et al., 2008, 2011), and 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.

In addition to its possible role in downregulating RT, the EhLINE1 antisense transcript could also serve as a long noncoding RNA which may have been co-opted by the host for other gene regulatory functions. This intriguing possibility needs to be tested. The involvement of LINEs in attenuating the expression of selected host genes has been documented. Chimeric transcripts driven by the ASP in L1 are believed to affect as many as 4% of all human genes, and may be important in modulating host gene expression (Criscione et al., 2016). In *E. histolytica* it remains to be seen whether some of the products of EhLINE1, like the constitutively expressed ORF1 polypeptide and the 1.5 kb antisense RT transcript have roles in cellular physiology other than those in retrotransposition.

5. Conclusion

We have provided a detailed account of the transcription status of individual EhLINE1 copies. 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. It remains to be seen whether sense transcripts corresponding to the functional protein domains are translatable under certain conditions and could mobilize RNAs *in trans*, and whether the antisense RT transcript could have regulatory functions other than in retrotransposition.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. However, the raw data used in this study has been submitted in public archive GEO, the GEO series accession number of which has already been provided. Please contact author for any other data set if needed.

Funding

This research was supported by J.C. Bose national fellowship, SERB, India; grant (SB37(1)/14/16/2017-BRNS) from Board of Research in Nuclear Sciences, India; and fellowship from Indian National Science Academy to SB; UGC BSR meritorious fellowship (DK), and CSIR fellowship to MA and SSS. The funders had no role in study design, data analysis, or writing of the manuscript.

Authors' contributions

DK, MA, SSS, PKM, AK, SB were involved in the study design. DK, MA, SSS conducted the study. DK, MA, AB, SB analyzed and interpreted the data. DK, MA, SB drafted the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.plasmid.2021.102560.

References

Alexandrova, E.A., Olovnikov, I.A., Malakhova, G.V., Zabolotneva, A.A., Suntsova, M.V., Dmitriev, S.E., Buzdin, A.A., 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.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/ doi:citeulike-article-id:11583827.

Bailey, T.L., Gribskov, M., 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. https://doi.org/10.1093/bioinformatics/14.1.48.

Bakre, A.A., Rawal, K., Ramaswamy, R., Bhattacharya, A., Bhattacharya, S., 2005. The LINEs and SINEs of *Entamoeba histolytica*: comparative analysis and genomic distribution. Exp. Parasitol. https://doi.org/10.1016/j.exppara.2005.02.009.

Belancio, V.P., Roy-Engel, A.M., Pochampally, R.R., Deininger, P., 2010. Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res. https://doi.org/10.1093/nar/gkq132.

Benihoud, K., Bonardelle, D., Soual-Hoebeke, E., Durand-Gasselin, I., Emilie, D., Kiger, N., Bobé, P., 2002. Unusual expression of LINE-1 transposable element in the MRL autoimmune lymphoproliferative syndrome-prone strain. Oncogene. https://doi.org/10.1038/sj.onc.1205730.

Branciforte, D., Martin, S.L., 1994. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. Mol. Cell. Biol. https://doi.org/10.1128/mcb.14.4.2584.

Bringaud, F., Müller, M., Cerqueira, G.C., Smith, M., Rochette, A., El-Sayed, N.M.A., Papadopoulou, 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.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Morant, J.V., Kazazian, H.H., 2003. Hot L1s account for the bulk of retrotransposition in the

human population. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.0831042100.
Bruchhaus, I., Leippe, M., Lioutas, C., Tannich, E., 1993. Unusual gene organization in the protozoan parasite *Entamoeba histolytica*. DNA Cell Biol. https://doi.org/10.1089/dna.1993.12.925.
Chaboissier, M.C., Busseau, I., Prosser, J., Finnegan, D.J., Bucheton, A., 1990. Identification of a potential RNA intermediate for transposition of the LINE-like element I factor in *Drosophila melanogaster*. EMBO J. https://doi.org/10.1002/j.1460-2075.1990.tb07566.x.
Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., Gage, F.H., 2009. L1 retrotransposition in human neural progenitor cells. Nature. https://doi.org/10.1038/nature08248.
Criscione, S.W., Theodosakis, N., Micevic, G., Cornish, T.C., Burns, K.H., Neretti, N., Rodić, N., 2016. Genome-wide characterization of human L1 antisense promoter-driven transcripts. BMC Genomics. https://doi.org/10.1186/s12864-016-2800-5.
Deininger, P.L., Batzer, M.A., 1999. Alu repeats and human disease. Mol. Genet. Metab. https://doi.org/10.1006/mgme.1999.2864.
Deininger, P., Morales, M.E., White, T.B., Baddoo, M., Hedges, D.J., Servant, G., Srivastav, S., Smither, M.E., Concha, M., DeHaro, D.L., Flemington, E.K., Belancio, V. P., 2017. A comprehensive approach to expression of L1 loci. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw1067.
Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C.N., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J., Moore, L., Hunter, T., Saghatelian, A., Gage, F.H., 2015. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell. https://doi.org/10.1016/j.cell.2015.09.025.
Diamond, L.S., Harlow, D.R., Cunnick, C.C., 1978. A new medium for the axenic cultivation of *Entamoeba histolytica* and other entamoeba. Trans. R. Soc. Trop. Med. Hyg. https://doi.org/10.1016/0035-9203(78)90144-X.
Dudley, J.P., 1987. Discrete high molecular weight RNA transcribed from the long interspersed repetitive element limd. Nucleic Acids Res. https://doi.org/10.1093/nar/15.6.2581.
Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. https://doi.org/10.1093/nar/30.1.207.
Eickbush, T.H., Malik, H.S., 2014. Origins and evolution of retrotransposons. Mobile DNA II. https://doi.org/10.1128/9781555817954.ch49.
Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Carninci, P., 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. https://doi.org/10.1038/ng.368.
Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., Bairoch, A., 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. https://doi.org/10.1093/nar/gkg563.
Gaurav, A.K., Kumar, J., Agrahari, M., Bhattacharya, A., Yadav, V.P., Bhattacharya, S., 2017. Functionally conserved RNA-binding and protein-protein interaction properties of LINE-ORF1p in an ancient clade of non-LTR retrotransposons of *Entamoeba histolytica*. Mol. Biochem. Parasitol. https://doi.org/10.1016/j.molbiopara.2016.11.004.
Goodier, J.L., 2016. Restricting retrotransposons: a review. Mob. DNA. https://doi.org/10.1186/s13100-016-0070-z.
Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr064.
Gupta, R., Dewan, I., Bharti, R., Bhattacharya, A., 2012. Differential expression analysis for RNA-Seq data. ISRN Bioinf. https://doi.org/10.5402/2012/817508.
Hamann, L., Buß, H., Tannich, E., 1997. Tetracycline-controlled gene expression in *Entamoeba histolytica*. Mol. Biochem. Parasitol. https://doi.org/10.1016/S0166-6851(96)02771-5.
Han, J.S., 2010. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob. DNA. https://doi.org/10.1186/1759-8753-1-15.
Hastings, K.E.M., 2005. SL trans-splicing: easy come or easy go? Trends Genet. https://doi.org/10.1016/j.tig.2005.02.005.
Heras, S.R., López, M.C., Olivares, M., Thomas, M.C., 2007. The L1Tc non-LTR retrotransposon of *Trypanosoma cruzi* contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl1137.
Hon, C.C., Weber, C., Sismeiro, O., Proux, C., Koutero, M., Deloger, M., Das, S., Agrahari, M., Dillies, M.A., Jagla, B., Coppee, J.Y., Bhattacharya, A., Guillen, N., 2013. Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1271.
Huang, C.R.L., Burns, K.H., Boeke, J.D., 2012. Active transposition in genomes. Annu. Rev. Genet. https://doi.org/10.1146/annurev-genet-110711-155616.
Hughes, T.E., Langdale, J.A., Kelly, S., 2014. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome Res. https://doi.org/10.1101/gr.172684.114.
Illumina, 2013. TruSeq® Stranded Total RNA Sample Preparation Guide. Illumina Technical Manuals. RS-200-9002DOC.
Iwasaki, Y.W., Siomi, M.C., Siomi, H., 2015. PIWI-interacting RNA: its biogenesis and functions. Annu. Rev. Biochem. https://doi.org/10.1146/annurev-biochem-060614-034258.
Jin, Y., Tam, O.H., Paniagua, E., Hammell, M., 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv422.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. https://doi.org/10.1093/molbev/mst010.
Kaul, T., Morales, M.E., Sartor, A.O., Belancio, V.P., Deininger, P., 2020. Comparative analysis on the expression of L1 loci using various RNA-Seq preparations. Mob. DNA. https://doi.org/10.1186/s13100-019-0194-z.
Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science. https://doi.org/10.1126/science.1089670.
Khadgi, B.B., Govindaraju, A., Christensen, S.M., 2019. Completion of LINE integration involves an open '4-way' branched DNA intermediate. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz673.
Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. https://doi.org/10.1038/s41587-019-0201-4.
Kimpton, C.P., Gill, P., Walton, A., Urquhart, A., Millican, E.S., Adams, M., 1993. Automated DMA profiling employing multiplex amplification of short tandem repeat loci. Genome Res. https://doi.org/10.1101/gr.3.1.13.
Kumari, V., Iyer, L.R., Roy, R., Bhargava, V., Panda, S., Paul, J., Verweij, J.J., Clark, C.G., Bhattacharya, A., Bhattacharya, S., 2013. Genomic distribution of SINEs in *Entamoeba histolytica* strains: implication for genotyping. BMC Genomics. https://doi.org/10.1186/1471-2164-14-432.
Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D.C., Moore, L., Nakashima, K., Asashima, M., Gage, F.H., 2009. Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat. Neurosci. https://doi.org/10.1038/nn.2360.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., Morgan, M.J., 2001. Initial sequencing and analysis of the human genome. Nature. https://doi.org/10.1038/35057062.
Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. https://doi.org/10.1186/1471-2105-12-323.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp352.
Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., Dewey, C.N., 2010. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp692.
López-Camarillo, C., Orozco, E., Marchat, L.A., 2005. *Entamoeba histolytica*: comparative genomics of the pre-mRNA 3′ end processing machinery. Exp. Parasitol. https://doi.org/10.1016/j.exppara.2005.02.024.
Lorenzi, H., Thiagarajan, M., Haas, B., Wortman, J., Hall, N., Caler, E., 2008. Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. BMC Genomics. https://doi.org/10.1186/1471-2164-9-595.
Macías, F., López, M.C., Thomas, M.C., 2016. The Trypanosomatid Pr77-hallmark contains a downstream core promoter element essential for transcription activity of the *Trypanosoma cruzi* L1Tc retrotransposon. BMC Genomics. https://doi.org/10.1186/s12864-016-2427-6.
Mandal, P.K., 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.
Martin, S.L., 1991. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol. Cell. Biol. https://doi.org/10.1128/MCB.11.9.4804.
Matsuda, E., Garfinkel, D.J., 2009. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. Proc. Natl. Acad. Sci. U. S. A. 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 li transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res. https://doi.org/10.1093/nar/20.12.3139.
Mizrokhi, L.J., Georgieva, S.G., Ilyin, Y.V., 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/S0092-8674(88)80013-8.
Naiyer, S., Kaur, D., Ahamad, J., Singh, S.S., Singh, Y.P., Thakur, V., Bhattacharya, A., Bhattacharya, S., 2019. Transcriptomic analysis reveals novel downstream regulatory motifs and highly transcribed virulence factor genes of *Entamoeba histolytica*. BMC Genomics. https://doi.org/10.1186/s12864-019-5570-z.
Ostertag, E.M., Kazazian Jr., H.H., 2001. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. https://doi.org/10.1146/annurev.genet.35.102401.091032.
Packer, A.I., Manova, K., Bachvarova, R.F., 1993. A discrete LINE-1 transcript in mouse blastocysts. Dev. Biol. https://doi.org/10.1006/dbio.1993.1133.
Perepelitsa-Belancio, V., Deininger, P., 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat. Genet. https://doi.org/10.1038/ng1269.
Petter, R., Rozenblatt, S., Nuchamowitz, Y., Mirelman, D., 1992. Linkage between actin and ribosomal protein L21 genes in *Entamoeba histolytica*. Mol. Biochem. Parasitol. https://doi.org/10.1016/0166-6851(92)90182-J.
Philippe, C., Vargas-Landin, D.B., Doucet, A.J., Van Essen, D., Vera-Otarola, J., Kuciak, M., Corbin, A., Nigumann, P., Cristofari, G., 2016. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. ELife. https://doi.org/10.7554/eLife.13926.
Pierce, J.C., Kong, D., Masker, W., 1991. The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. Nucleic Acids Res. https://doi.org/10.1093/nar/19.14.3901.

Purdy, J.E., Pho, L.T., Mann, B.J., Petri Jr., W.A., 1996. Upstream regulatory elements controlling expression of the *Entamoeba histolytica* lectin. Mol. Biochem. Parasitol. 78 (1-2), 91–103. https://doi.org/10.1016/s0166-6851(96)02614-x.

Rangwala, S.H., Zhang, L., Kazazian, H.H., 2009. Many LINE1 elements contribute to the transcriptome of human somatic cells. Genome Biol. https://doi.org/10.1186/gb-2009-10-9-r100.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., Kazazian, H.H., 1997. Many human L1 elements are capable of retrotransposition. Nat. Genet. https://doi.org/10.1038/ng0597-37.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., Margolet, L., 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics. https://doi.org/10.1016/0888-7543(87)90003-6.

Seol, J.H., Shim, E.Y., Lee, S.E., 2018. Microhomology-mediated end joining: good, bad and ugly. In: Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. https://doi.org/10.1016/j.mrfmmm.2017.07.002.

Sfeir, A., Symington, L.S., 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? Trends Biochem. Sci. https://doi.org/10.1016/j.tibs.2015.08.006.

Sharma, S., Banyal, N., Singh, M., Mandal, A.K., 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.

Shrimal, S., Bhattacharya, S., Bhattacharya, A., 2010. Serum-dependent selective expression of EhTMKB1-9, a member of *Entamoeba histolytica* B1 family of transmembrane kinases. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1000929.

Singh, S.S., Naiyer, S., Bharadwaj, R., Kumar, A., Singh, Y.P., Ray, A.K., Subbarao, N., Bhattacharya, A., Bhattacharya, S., 2018. Stress-induced nuclear depletion of *Entamoeba histolytica* 3′-5′ exoribonuclease EhRrp6 and its role in growth and erythrophagocytosis. J. Biol. Chem. https://doi.org/10.1074/jbc.RA118.004632.

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.

Stothard, P., 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques. https://doi.org/10.2144/00286ir01.

Swergold, G.D., 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol. Cell. Biol. https://doi.org/10.1128/MCB.10.12.6718. Updated.

Trelogan, S.A., Martin, S.L., 1995. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.92.5.1520.

Trinh, T.Q., Sinden, R.R., 1993. The influence of primary and secondary DNA structure in deletion and duplication between direct repeats in *Escherichia coli*. Genetics. 134 (2), 409–422.

Vesteg, M., Hadariová, L., Horváth, A., Estraño, C.E., Schwartzbach, S.D., Krajčovič, J., 2019. Comparative molecular cell biology of phototrophic euglenids and parasitic trypanosomatids sheds light on the ancestor of Euglenozoa. Biol. Rev. https://doi.org/10.1111/brv.12523.

Willhoeft, U., Hamann, L., Tannich, E., 1999. A DNA sequence corresponding to the gene encoding cysteine proteinase 5 in *Entamoeba histolytica* is present and positionally conserved but highly degenerated in *Entamoeba dispar*. Infect. Immun. 67 (11), 5925–5929. https://doi.org/10.1128/IAI.67.11.5925-5929.1999.

Wissing, S., Mūoz-lopez, M., Macia, A., Yang, Z., Montano, M., Collins, W., Garcia-perez, J.L., Moran, J.V., Greene, W.C., 2012. Reprogramming somatic cells into ips cells activates line-1 retroelement mobility. Hum. Mol. Genet. https://doi.org/10.1093/hmg/ddr455.

Yadav, V.P., Mandal, P.K., Rao, D.N., 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, V.P., Mandal, P.K., 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, H.S., Eickbush, T.H., 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.96.14.7847.

Yang, W.R., Ardeljan, D., Pacyna, C.N., Payer, L.M., Burns, K.H., 2019. SQuIRE reveals locus-specific regulation of interspersed repeat expression. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1301.

Zamorano, A., López-Camarillo, C., Orozco, E., Weber, C., Guillen, N., Marchat, L.A., 2008. In silico analysis of EST and genomic sequences allowed the prediction of cis-regulatory elements for *Entamoeba histolytica* mRNA polyadenylation. Comput. Biol. Chem. https://doi.org/10.1016/j.compbiolchem.2008.03.019.

Zhang, H., Ehrenkaufer, G.M., Pompey, J.M., Hackney, J.A., 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., Alramini, 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.