Available online at www.sciencedirect.com
SCIENCE DIRECT®
Experimental Parasitology 110 (2005) 207–213
Experimental Parasitology
www.elsevier.com/locate/yexpr

The LINEs and SINEs of *Entamoeba histolytica*: Comparative analysis and genomic distribution

Abhijeet A. Bakre [a], Kamal Rawal [b], Ram Ramaswamy [b,d,1], Alok Bhattacharya [b,c], Sudha Bhattacharya [a,*]

[a] *School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India*
[b] *Bioinformatics Centre, School of Information Technology, Jawaharlal Nehru University, New Delhi, India*
[c] *School of Life Sciences, Jawaharlal Nehru University, New Delhi, India*
[d] *School of Physical Sciences, Jawaharlal Nehru University, New Delhi, India*

Received 31 January 2005; received in revised form 31 January 2005; accepted 15 February 2005
Available online 23 March 2005

Abstract

Autonomous non-long terminal repeat retrotransposons are commonly referred to as long interspersed elements (LINEs). Short non-autonomous elements that borrow the LINE machinery are called SINES. The *Entamoeba histolytica* genome contains three classes of LINEs and SINEs. Together the EhLINEs/SINEs account for about 6% of the genome. The recognizable functional domains in all three EhLINEs included reverse transcriptase and endonuclease. A novel feature was the presence of two types of members—some with a single long ORF (less frequent) and some with two ORFs (more frequent) in both EhLINE1 and 2. The two ORFs were generated by conserved changes leading to stop codon. Computational analysis of the immediate flanking sequences for each element showed that they inserted in AT-rich sequences, with a preponderance of Ts in the upstream site. The elements were very frequently located close to protein-coding genes and other EhLINEs/SINEs. The possible influence of these elements on expression of neighboring genes needs to be determined.

Index Descriptors and Abbreviations: EhLINEs; EhSINEs; Non-LTR retrotransposons; Genomic distribution; *E. histolytica*; GSS, genome survey sequences; LINE, long interspersed nuclear element; SINE, short interspersed nuclear element; ORF, open reading frame; nt, nucleotide; aa, amino acid; RT, reverse transcriptase; EN, endonuclease

Keywords: EhLINEs; EhSINEs; Non-LTR retrotransposons; Genomic distribution; *E. histolytica*

1. Introduction

Repetitive elements have shaped genomic organization and can influence gene expression (Landry et al., 2001). Repetitive DNA exists to varying extent in the genomes of protozoan parasites (Bhattacharya et al., 2002; Wickstead et al., 2003), including *Entamoeba histolytica*. This parasite is the causative agent of amoebiasis, a highly prevalent disease in developing countries. It has a 23 Mb genome consisting of 14–17 linear chromosomes and numerous episomes, the most abundant of which is the 24.5 kb ribosomal DNA circle (Bhattacharya et al., 2000; Willhoeft and Tannich, 1999). *E. histolytica* is not only a clinically important organism but also occupies a unique niche in evolution.

Eukaryotic genomes are home to various types of transposons (Craig, 2002), of which the non-long terminal repeat (LTR) retrotransposons are abundantly found in *E. histolytica*. Autonomous Non-LTR elements

* Corresponding author.
E-mail address: sb@mail.jnu.ac.in (S. Bhattacharya).
[1] Present address: Centre for Systems Biology, School of Natural Sciences, Institute for Advanced Study, Princeton NJ 08540, USA.

0014-4894/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.exppara.2005.02.009

encoding their own retrotransposition machinery are commonly referred to as long interspersed elements (LINEs). Short non-autonomous elements that borrow this machinery for propagation are called short interspersed elements (SINES). LINEs and SINEs profoundly influence the host genome via a multitude of mechanisms (Ostertag and Kazazian, 2001). They may affect gene expression by providing alternative promoters, splicing and polyadenylation sites, and by heterochromatinization. In addition, SINEs can also work as stress sensors in the cell (Kimura et al., 2001). Studies with repetitive DNAs of *E. histolytica* revealed a 4.8 kb element, part of which had a very close match with reverse transcriptase (RT) of non-LTR retrotransposons (Sharma et al., 2001). Another repetitive and highly transcribed 0.55 kb element was discovered, which lacked an open reading frame (ORF) (Cruz-Reyes et al., 1995; Willhoeft et al., 1999). This element shared a 70 nt sequence at the 3′-end with the 4.8 kb element, and the two were proposed to be a LINE/SINE pair (Bhattacharya et al., 2002; Willhoeft et al., 2002). Analysis of the *E. histolytica* genome sequence database showed the existence of multiple families of autonomous and non-autonomous non-LTR elements, now designated EhLINEs and EhSINEs (Van Dellen et al., 2002). The discovery of a EhLINE-encoded endonuclease (EN) activity which could nick a natural target site of EhSINE insertion, provided evidence that EhSINE1 could utilize the EhLINE1 machinery for its own transposition (Mandal et al., 2004).

In this article, we report the comparative characterization of three families of LINEs and SINEs in the *E. histolytica* genome with respect to their sequence organization and genomic distribution.

2. Materials and methods

2.1. Identification and assembly of the three families of EhLINEs and EhSINEs

Entamoeba histolytica GSS sequences at NCBI were used for initial element assembly. EhLINE1, 2, and 3 were assembled using standard procedures. IE (AF126955) (Cruz-Reyes et al., 1995) was renamed as EhSINE1 after a consensus sequence was derived from multiple alignment of GSS clones using majority rule. EhSINE2 shares a stretch of ~70 nt with EhSINE1 at the 5′ end and was thus identified. EhSINE3 was constructed as an *E. histolytica* homologue for the abundant polyadenylated transcript UEE from *Entamoeba dispar* (Sharma et al., 1999) using GSS clones BH167278 and AZ545188. Assembled elements were checked for exact copies in the final *E. histolytica* genome. Full-length and truncated elements were extracted from the *E. histolytica* final genome assembly at NCBI, which represents the complete non-redundant haploid genome of *E. histolytica* (www.ncbi.nlm.nih.gov). All sequence analysis including pair wise alignment (BLAST) and multiple alignments (CLUSTAL W ver.1.8) was carried out using the BioEdit suite of programs locally or at NCBI.

2.2. Mining of truncated and full-length EhLINEs and EhSINEs

Consensus sequence for each EhLINE and EhSINE was used in a BLAST search against the *E. histolytica* genome database to identify full-length and truncated copies. All significant hits ($p < 1E\text{-}8$) and a 2 kb region flanking them were retrieved. Elements matching both the 5′ and 3′ ends were designated complete, while those, which did not match either the start or end of the element consensus sequence, were termed truncated. To remove redundant hits, all 5′ flanking sequences were compared via pairwise global alignment, and pairs with percentage identity score exceeding 90% in the flanking region were deemed to be redundant. ELEANALYZER, a software tool was developed and used for this purpose (manuscript in preparation). Flanking regions were further inspected for the presence of coding sequences through BLASTX searches against the nr database. Flanking regions of each copy were examined for occurrence of another instance of an element. The coding potential of all full-length copies was also determined for each EhLINE family.

2.3. Interfamily and intrafamily conservation analysis of EhLINEs and EhSINEs

Interfamily conservation was computed at both nucleotide and amino acid level. Full-length consensus elements were used for overall identity at nucleotide level. Nucleotide and amino acid regions corresponding to ORF-1, ORF-2, RT, and EN were used as queries against the database. Full-length hits to queries above threshold identity (identity >80%) were used to calculate mean intra-family identity. For intra-family domain identity a matrix of conserved amino acids believed to be functionally important in the RT and EN domain was constructed and used for computing identity (Moran and Gilbert, 2002).

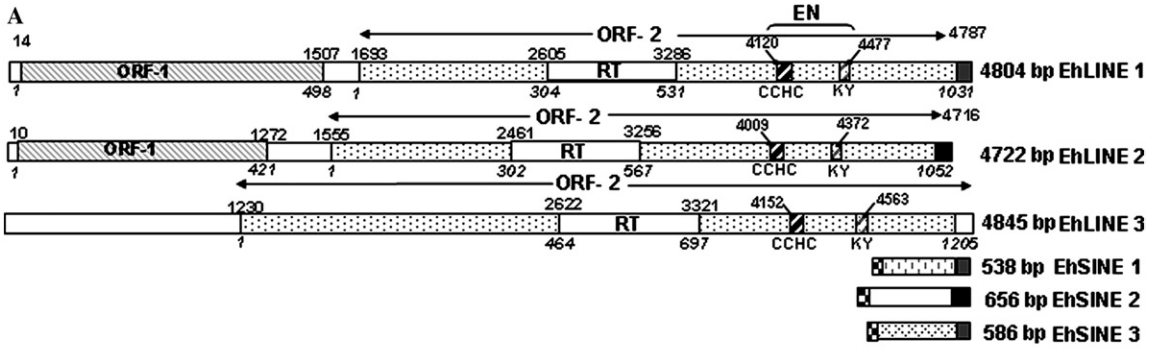
3. Results

3.1. Comparative sequence analysis of EhLINEs and EhSINEs

The *E. histolytica* genome contains three classes of LINEs (EhLINE1, 2, and 3) and two classes of SINEs (EhSINE1 and 2) with a third class of EhSINE that appears to be present in a single copy. The size and copy

number of each element as deduced from the *E. histolytica* genome sequence submitted with NCBI is shown in Fig. 1. Most copies of each element are truncated at the 5′- or 3′-end or at both ends. Of the full-length copies none was found to contain a complete ORF, due to many point mutations. A consensus sequence of each EhLINE, with a complete ORF was reconstructed manually by selecting the most common nucleotide at each position. Analysis of the consensus sequence showed that EhLINE1 had a length of 4804 bp and consisted of two easily identifiable functional domains—the RT (nt 2605–3286), and the EN (nt 4120–4477). The RT domain showed the closest match with RTs encoded by the R4 clade of non-LTR retrotransposons, most notably the R4 element of *Ascaris lumbricoides* and the Dong element of *Bombyx mori*. The 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 elements (Bhattacharya et al., 2002; Van Dellen et al., 2002). The N-terminal one-third of the element encoded a polypeptide (ORF-1) with some matches with proteins containing coiled coil domains. Similar analysis of EhLINE2 and 3 is presented in Fig. 1. These were present in fewer copies than EhLINE1, but the overall sequence organization was very similar. The RT and EN domains were well conserved in EhLINE2 and 3. However, the consensus sequence reconstructed for EhLINE3 did not have ORF-1. This may be due to the accumulation of too many mutations in this part of the element. EhLINE3 was not only present in the fewest copies amongst the three EhLINEs, most of the copies (75%) were truncated at both ends. By virtue of the sequence identity at the 3′-end between EhLINEs and EhSINEs (87% identity in a 73 nt stretch between EhLINE1 and EhSINE1, and 76% identity in a 84 nt stretch between EhLINE2 and EhSINE2) (Fig. 1), EhSINE1 and 2 are thought to be non-autonomous elements (Kajikawa and Okada, 2002). Only one copy of EhSINE3 was found.

None of the full-length copies of EhLINEs in the database had complete ORFs. Interfamily sequence comparison between all the members of EhLINEs showed that although the overall sequence identity was low, the functionally important amino acid residues in the RT and EN



B

EhLINE 1 EN	*809* C (2) C (8) H (4) C (9) KHD (25) E (26) R P D (14) D (15) K (3) Y	*928*
EhLINE 2 EN	*818* C (2) C (8) H (4) C (9) KHD (25) P (26) R P D (14) D (15) K (3) Y	*939*
EhLINE 3 EN	*974* C (2) C (8) H (4) C (9) KHD (26) E (44) R I D (14) D (14) K (3) Y	*1111*

C

Element	Copy no	Full length	Only 5' end intact	Only 3'end intact	Neither end intact
EhLINE1	409	49 (12%)	95 (23%)	121 (30%)	144 (35%)
EhLINE2	290	28 (10%)	56 (19%)	92 (32%)	114 (39%)
EhLINE3	52	3 (6%)	3 (6%)	7 (13%)	39 (75%)
EhSINE1	272	81 (30%)	18 (7%)	88 (32%)	85 (31%)
EhSINE2	117	62 (53%)	16 (14%)	8 (7%)	31 (26%)
EhSINE3	Only single copy exists				

Fig. 1. Sequence organization and genomic abundance of full-length and truncated copies of EhLINEs and EhSINEs. (A) Organization of full-length EhLINE1, 2, and 3 and EhSINE1, 2, and 3. Consensus sequence of each EhLINE family derived by comparative analysis of all the entries in the data base were used to mark the ORFs and other features, including the 5′- and 3′-ends of each family. The RT and EN domains in ORF2 are indicated. Within the EN domain the CCHC and KY conserved motifs are also marked. The elements are drawn to scale. Numbers on top of each LINE family denote nucleotide positions while numbers below the LINE (in italics) denote amino acids in the ORF. Regions identical between EhLINEs and EhSINEs at their 3′-ends are shown by similar shading. The stretch of sequence homology at the 5′-ends of EhSINEs is also indicated. (B) Amino acid alignment of the EN domain of EhLINEs. The highly conserved amino acids thought to be functionally important are shaded. Numbers in bracket correspond to the number of amino acids found at those positions. Numbers in italics indicate the amino acid position at the start and end of each EN domain. (C) Genomic abundance of full-length and truncated copies of EhLINEs and EhSINEs. Copy number was calculated as given in Section 2.

Table 1
Inter- and Intrafamily conservation between EhLINEs and EhSINEs

Interfamily	Nucleotide identity (%)		Identity in the ORF and domain regions (%)							
	Full-length copies		ORF-1		ORF-2					
					Overall		RT		EN	
			nt	aa	nt	aa	nt	aa	nt	aa
EhLINE1 vs EhLINE2	52		45	15	53	35	50	76	44	85
EhLINE1 vs EhLINE3	56		NA		45	34	45	80	48	92
EhLINE2 vs EhLINE3	52		NA		42	27	41	51	52	85
EhSINE1 vs EhSINE2	40		No ORF							
EhSINE1 vs EhSINE3	40		No ORF							
EhSINE2 vs EhSINE3	38		No ORF							
Intrafamily	Full-length	Truncated copies								
EhLINE 1	95	91	96	94	94	84	94	93	93	94
EhLINE 2	93	93	94	87	94	85	94	93	93	91
EhLINE 3	94	90	ORF-1 not found		95	87	95	81	96	91
EhSINE 1	94	87	No ORF							
EhSINE 2	97	90	No ORF							
EhSINE 3	NA		No ORF							

Inter- and Intrafamily identities between full elements, ORFs, and domains were calculated as given in Section 2.
RT, reverse transcriptase; EN, Endonuclease; nt, nucleotide; aa, amino acid; NA, Not applicable.

domains were well conserved (Table 1). The high overall amino acid identity of ORF-1 in the intrafamily analysis indicates that this protein performs a conserved function. Although most copies of EhLINE3 were truncated (Fig. 1), intrafamily sequence comparison showed that the nt sequence identity of this family was comparable with the values for EhLINEs1 and 2 (Table 1).

3.2. EhLINEs may either have a single ORF or two ORFs

The reconstructed copy of a full-length EhLINE1 with complete ORFs showed the presence of two long ORFs (nt 14–1507, and 1693–4787). However, another reconstructed copy of EhLINE1 was reported as having a single ORF (Van Dellen et al., 2002). Comparison of the two copies showed that a 5 nt sequence (AAGCA) was duplicated at position 1442 in the element containing two ORFs. This resulted in a stop codon at position 1507 (Fig. 2). The putative start of the second ORF was assigned at position 1693. The deduced amino acid sequence of the proteins encoded by the one-ORF and two-ORF elements was identical, except for the missing amino acids between the two ORFs. These amino acids may not be functionally important and had no match with known sequences in the database. When all EhLINE1 sequences in the database were searched for presence or absence of the 5 nt duplication, about 80% were found to contain the duplication, resulting in two ORFs. To see whether one or the other of these elements was preferentially transcribed, RT-PCR was performed with total *E. histolytica* RNA to amplify a 1.3 kb fragment containing the region between the two ORFs. The amplified cDNA was cloned and several independent clones were sequenced. Both types of sequence—with and without the 5 nt duplication were present, showing a lack of transcriptional bias of the two types of EhLINE1.

EhLINE2 also had some members with a single ORF (nt 10–4716) and some with two ORFs (nt 10–1272, and 1555–4716) (Fig. 2). The two-ORF element contained a deletion of two nt (CG) at position 1249, resulting in a

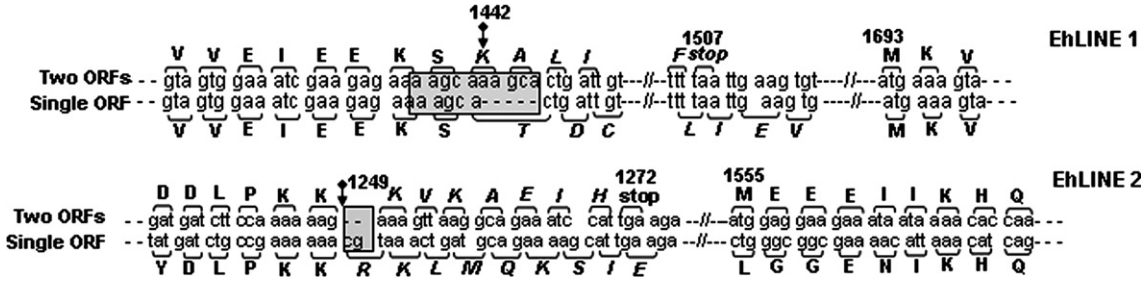


Fig. 2. Nucleotide changes leading to copies with either one or two ORFs in EhLINEs. In EhLINE1 a 5 bp duplication (AAGCA) at nt position 1442 leads to copies with two ORFs due to introduction of a stop codon at position 1507. The putative start of ORF2 could be at the AUG at position 1693. In EhLINE2, a 2 bp deletion (CG) at nt position 1249 leads to copies with two ORFs due to stop codon at position 1272. The putative start of ORF2 could be at the nt position 1555. Amino acids corresponding to the codons are indicated.

stop codon at position 1272. Alignment of GSS clones in the region between the two ORFs showed that majority of clones (92%) had this deletion resulting in two ORFs.

3.3. Genomic distribution of EhLINEs and EhSINES

Previous Southern hybridization studies of PFGE separated chromosomes of *E. histolytica* with EhLINE1 and EhSINE1 probes (Bagchi et al., 1999) showed that these elements reside on all chromosomal bands, do not seem to be telomeric, and might be dispersed in the *E. histolytica* genome. Computational analysis of the immediate upstream and downstream flanking sequences for each element supported this view, since no conserved sequences could be found at the sites of insertion of any of the elements. However, all the elements 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. 3). A 2 kb region surrounding each element was searched for the presence of protein-coding genes and other instances of EhLINEs/SINEs. The analysis for the 2 kb region upstream is presented. The same results were obtained for the downstream region and there was substantial overlap in the data. Only in 20% cases, neither a gene nor an element was found within 2 kb. The chance of finding a gene was about 50% greater than that of finding another element within the 2 kb. The pattern of occurrence of genes or elements in the vicinity of EhLINEs/SINEs is depicted in Fig. 4. In general, 50% of the time, when a gene was found near an element, it was present within the first 0.5 kb, whereas an element was found within the first 0.1 kb. Such close proximity of elements to one another could be due to clustering of sequences that serve as favorable target sites for insertion. Elements were found in both orientations with respect to each other. No clear bias of any pairs of elements occurring near each other could be discerned. Amongst the genes present near the elements, most (about 63%) were found to be hypothetical. Of the genes that gave a match in the database the most common class was that of protein kinases. Other genes found in the vicinity were GTPases, heat shock proteins, and BspA. House keeping genes were rarely found. No instance was encountered of an element inserted within a gene.

4. Discussion

The EhLINEs/SINEs together account for 6% of the *E. histolytica* genome as deduced from data base analysis. The other types of transposable elements—DNA transposons and LTR-retrotransposons seem to be absent in *E. histolytica*. Various types of non-LTR retrotransposons are encountered in living organisms. These differ from one another in several ways, including the kind of endonuclease encoded by the element (restriction enzyme-like or apurinic endonuclease), the number of ORFs, the relative arrangement of functional domains, and the specificity of target site for insertion (sequence specific versus dispersed). In *E. histolytica*, the three different families of EhLINEs share all these basic features with one another, and may

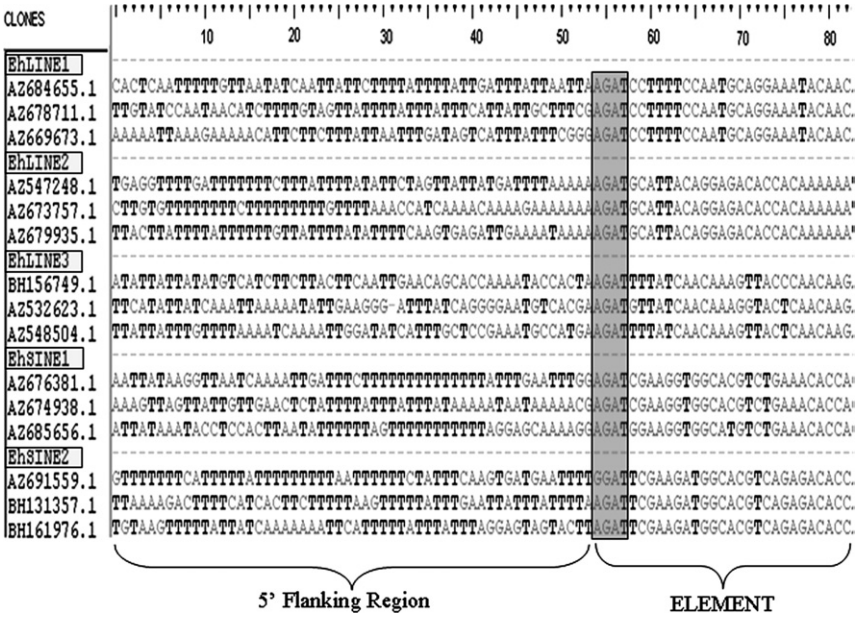


Fig. 3. T stretches in the 5′ flanking sequences of EhLINEs/SINEs. Three examples of each element are shown. The same pattern was observed in most copies. The first four nucleotides (AGAT) conserved in all elements are shaded. Accession number of each database entry is indicated on the left.

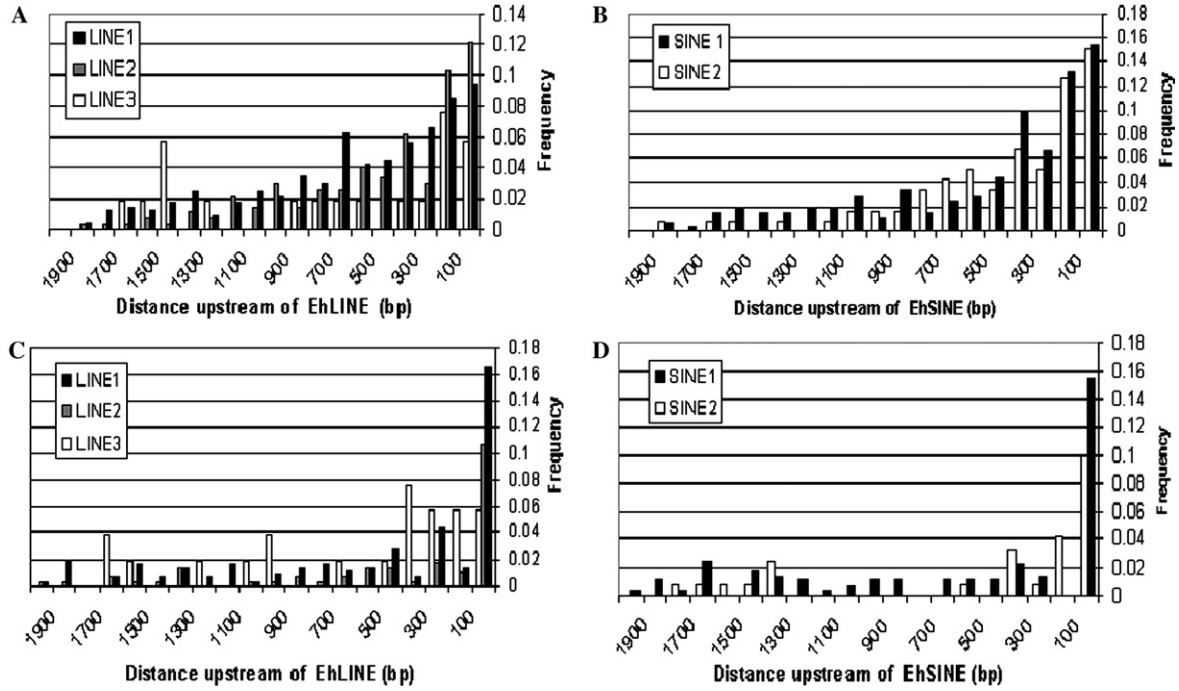


Fig. 4. Frequency of occurrence of protein coding genes (A,B) or another instance of EhLINE/EhSINE (C,D) in a 2 kb region upstream of EhLINEs and EhSINEs. The frequency was computed as number of instances of an element associated with a gene or EhLINE/EhSINE at the indicated distances (within 2 kb upstream) divided by total number of hits of that type of element in the genome. Total number of hits includes all full-length and truncated copies of each element as shown in Fig. 1C.

have diverged from a common ancestor. All three have a restriction enzyme-like EN domain located 3′ of the RT domain. The target site sequences in which they insert all share the same features—namely an AT-rich sequence, with a T-rich stretch immediately upstream of the element. They are dispersed throughout the genome, frequently located near protein-coding genes. The insertion sites of EhSINEs also share the same features. Thus, in present day *E. histolytica* the transposable elements are all targeted to very similar genomic locations found mainly in intergenic regions. Whether the different EhLINEs/SINEs show subtle preference for certain genomic sites remains to be seen. Such preferences, if they exist, may be determined, in part, by the DNA sequences most readily nicked by the EN encoded by each EhLINE family. We are testing the substrate specificity of EhLINE1 EN (Mandal et al., 2004) and will be comparing it with other ENs. A novel observation in our analysis was the presence of two types of members- one with a single long ORF and one with two ORFs in both EhLINE1 and EhLINE2. Closely related non-LTR retrotransposons with either one or two ORFs have been reported. SLACS, in *Trypanosoma brucei* and CZAR in *Trypanosoma cruzi* have two ORFs while CRE1 and CRE2 of *Crithidia fasciculata* have a single ORF (Bhattacharya et al., 2002). However, the occurrence of one-ORF and two-ORF elements of the same type in a single organism as seen in *E. histolytica* has not been reported so far. From genome sequence data it was seen that the number of two-ORF elements was larger than the one-ORF elements for both EhLINE1 and EhLINE2. It is possible that two ORFs may impart greater stability and/or activity to the encoded polypeptides compared with a single large protein, and may have been selected for. This contention is supported by the very similar position of the stop codons in EhLINE1 and EhLINE2 leading to the generation of two ORFs. In both types of element the stop codon disrupted the single ORF at a distance of about one-thirds of the length from the N-terminus. The genome of *E. histolytica* shows extensive chromosome-length polymorphism amongst different strains (Bagchi et al., 1999; Willhoeft and Tannich, 1999). EhLINEs/SINEs, by virtue of their dispersed location in the genome, could be mediators of this polymorphism due to recombination, or DNA rearrangements associated with retrotransposition. The frequent location of EhLINEs/SINEs close to protein-coding genes leads to the possibility that these elements may influence gene-expression. Indeed, transcriptional silencing of the amoebapore gene has been demonstrated to be an epigenetic phenomenon that may be influenced by the EhSINE1 element located close by Bracha et al. (2003). In terms of pathogenesis of amoebiasis it would be important to determine the extent to which these elements may influence the phenotype of *E. histolytica* compared with the sibling, non-pathogenic species *Entamoeba dispar*.

Acknowledgments

This work was supported by a grant from Indian Council of Medical Research (ICMR) and University Grants Commission, India. A.A B. and K.R. are grateful to the Council of Scientific and Industrial Research (CSIR) and ICMR for fellowship, respectively.

References

Bagchi, A., Bhattacharya, A., Bhattacharya, S., 1999. Lack of a chromosomal copy of the circular rDNA plasmid of *Entamoeba histolytica*. International Journal of Parasitology 11, 1775–1783.

Bhattacharya, A., Satish, S., Bagchi, A., Bhattacharya, S., 2000. The genome of *Entamoeba histolytica*. International Journal of Parasitology 30, 401–410.

Bhattacharya, S., Bakre, A., Bhattacharya, A., 2002. Mobile genetic elements in protozoan parasites. Journal of Genetics 81, 73–86.

Bracha, R., Nuchamowitz, Y., Mirelman, D., 2003. Transcriptional silencing of an amoebapore gene in *Entamoeba histolytica*: molecular analysis and effect on pathogenicity. Eukaryotic Cell 2, 295–305.

Craig, N.L., 2002. Mobile DNA II. ASM Press, Washington, DC.

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

Kajikawa, M., Okada, N., 2002. LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell 111, 433–444.

Kimura, R.H., Choudary, PV., Stone, KK., Schmid, CW., 2001. Stress induction of Bm1 RNA in silkworm larvae: SINEs, an unusual class of stress genes. Cell Stress Chaperones 6, 263–272.

Landry, J.R., Medstrand, P., Mager, D.L., 2001. Repetitive elements in the 5′ untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. Genomics 76, 110–116.

Mandal, P., 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. Eukaryotic Cell 3, 170–179.

Moran, J.V., Gilbert, N., 2002. Mobile DNA II. ASM Press, Washington, DC.

Ostertag, E., Kazazian Jr., H.H., 2001. Biology of mammalian L1 retrotransposons. Annual Review of Genetics 35, 501–538.

Sharma, R., Azam, A., Bhattacharya, S., Bhattacharya, A., 1999. Identification of novel genes of non-pathogenic *Entamoeba dispar* by expressed sequence tag analysis. Molecular and Biochemical Parasitology 99, 279–285.

Sharma, R., Bagchi, A., Bhattacharya, S., Bhattacharya, A., 2001. Characterization of a retrotransposon-like repetitive DNA in *Entamoeba histolytica*. Molecular and Biochemical Parasitology 116, 45–53.

Van Dellen, K., Field, J., Wang, Z., Loftus, B., Samuelson, J., 2002. LINEs and SINE-like elements of the protist *Entamoeba histolytica*. Gene 297, 229–239.

Wickstead, B., Ersfeld, K., Gull, K., 2003. Repetitive elements in genomes of parasitic protozoa. Microbiology and Molecular Biology Reviews 67, 360–375.

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 150, 61–70.

Willhoeft, U., Tannich, E., 1999. The electrophoretic karyotype of *Entamoeba histolytica*. Molecular and Biochemical Parasitology 99, 41–53.

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*. Infection and Immunity 70, 6798–6804.