Expression analysis of *Entamoeba histolytica* Retrotransposons

Thesis submitted to
Jawaharlal Nehru University
for the award of

Doctor of Philosophy

MRIDULA AGRAHARI

SCHOOL OF ENVIRONMENTAL SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI-110067
INDIA
2017

SCHOOL OF ENVIRONMENTAL SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI 110067
INDIA

CERTIFICATE

The research work embodied in the thesis entitled **"Expression analysis of *Entamoeba histolytica* Retrotransposons"** has been carried out in School of Environmental Sciences, Jawaharlal Nehru University, New Delhi.

This work is original and has not been submitted so far, in part or in full, for award of any degree or diploma of any university.

Mridula

Mridula Agrahari
(Candidate)

Dr. Vijay Pal Yadav
(Supervisor)

Sudha Bhattacharya

Prof. Sudha Bhattacharya
(Co-supervisor)

Prof. S. Mukherjee
(Dean)

प्रो.सौमित्र मुखर्जी / Prof. S. Mukherjee
डीन / Dean
पर्यावरण विज्ञान संस्थान
School of Environmental Sciences
जवाहरलाल नेहरू विश्वविद्यालय,
Jawaharlal Nehru University
नई दिल्ली-110067/New Delhi-110067

To
The Almighty God

Acknowledgements

The process of earning a doctorate and writing a thesis is long and arduous and it is certainly not done single handily. Over the course of long and enjoyable period, I have been fortunate to receive help, guidance, and support from many people. This acknowledgement is the note of thanks for the support I received throughout this memorable period of my life.

I wish to pay my heartfelt gratitude to my mentor and guide Prof. Sudha Bhattacharya, whose stimulating suggestions and encouragement gave direction to my work. She has always been a constant source of inspiration and working under her guidance has been an enriching experience for me. Her unwavering enthusiasm for science kept me constantly engaged with the research and her personal generosity helped in making my time enjoyable in JNU. The internship opportunity I had with ma'am was a great chance for learning and professional development. Her challenges brought this work towards completion. I am sincerely grateful for her insight, patients, and her reviewing skills in helping me to structure and write the thesis dissertation. I really thank you, ma'am, for your support and care specially, when I was infected with Dengue fever. It was the most memorable moment for me when I saw u more like a guardian than a mentor.

I am equally thankful to Dr. Vijay Pal Yadav for accepting me as his student. Due to his cordial nature and friendly behavior, I was able to interact him which helped me to learn many valuable things. I highly admire his friendly and humble nature. I express my deepest thanks to him for his necessary advice and guidance which were extremely valuable for my research.

I am highly indebted to Prof. Alok Bhattacharya for finding out time and giving his precious and kind suggestions during lab meeting. Thank you, sir, for being a constant source of motivation and to broaden my horizon. I wish to thank him and ma'am for our numerous lab trips over the time, which has always been exciting and stress busting.

I am very grateful to Dr. Kausik Chakraborty for his much-appreciated suggestions and allowing me to work in his lab. I am thankful to Manish rai for helping me with experiments in his lab.

I would like to express my sincere thanks to all my teachers who put their faith in me and inspired me to do better.

Fellowship from UGC and CSIR is duly acknowledged.

I thank all the administrative staff and Dean of School of Environmental Sciences for their cooperation and facilities. I thank the CIF staffs of SES for making things available and their concern.

I sincerely thank Jeevan Ji, Manoj ji, Vikas, Subbu ji and Dinesh Ji for their help and arranging all the facilities to make life easier. Jeevan ji and Manoj ji appeared like the pillars of the lab, it is very difficult to run the lab in absence of either. Sometimes having lunch with them was really a healthy experience.

My appreciation extends to my laboratory colleagues. I thank my past SES lab mates as I shared their expertise with me very generously and learned a lot from them. With boundless love and appreciation, I would like to extend my heartfelt thanks to the people who helped me bring this study into reality. I am privileged for having Abhishek, Vandana, Nishant, Jitender, Sandeep, Ankita, Jamal, Amit, Ashwini, Sarah, Shashi, Shraddha, Devinder, YP, and Maneet as my colleges who have provided great company and cooperation in the lab. I really enjoyed the company of Jamal, Sarah, and Shashi when we use to go for food hunting to satisfy our taste buds. I am equally thankful to my SLS lab mates, Hafeez, Saima, Shahid, Arpita, Mrigya, Aruna, Ravi, Sabir, Pamchui, Shalini, Janhawi and Kapila for their cooperation, help and for good times during our lab trips. In particular, I would specially like to thank Shiteshu and Somlata for their generous support and guidance whenever needed. It was a pleasure working with them.

I want to say thanks to the JNU for letting me fulfill my dream of being a doctorate. Working as a Ph.D. student in JNU was a magnificent as well as a pleasant experience to me. In all these years, many people were instrumental directly or indirectly in shaping up my academic career. It was hardly possible for me to thrive in my doctoral work without the precious support of my friends Ranjeet, Neha, Ritu di, Rohit, Chandu, Arif, Madhav, Saurabh, Susheel, Bipasha, and Anshu.

Most importantly, thank you Ma and Pa for always just wanting the best for me. Your unconditional love, care, concern, and trust in me has made me a stronger person. I can never thank you enough for making your children the main priority in your lives. Without your love and constant support over the years, it would not have been possible for me to achieve my educational goals.

Above ground, I am indebted to my family. They have always been there for me and I am thankful for everything they have helped me achieve. Thanks, and love to my Mum and Dad including Didi, Jiju, Jyoti di and Ravi Bhai; all of whom have been there for at least part of the time and have been exceptionally patient throughout.

Finally, I acknowledge my husband and friend, Manish, who blessed me with a life of joy in the hours when the lab lights were off. He has been unfailingly supportive as I spent my time pursuing goals that took me away from him and family. Thank you, my friend, for helping me with proofreading and other technical problems I encountered along the way. Thesis writing would not have been possible without your unconditional support and love.

The skills and knowledge which I have gained throughout my practical training, I perceive as very valuable component and a big milestone in my future career development. I will strive to use gained skills and knowledge in the best possible way and will continue to work on their improvement.

Mridula

Abbreviations and Symbols

α	Alpha
γ	Gamma
µg	Microgram
µl	Microliter
µM	Micromolar
µCi	Microcurie
ATP	Adenosine triphosphate
bp	base pair
$CHCl_3$	Chloroform
Cl	Chloride
DEPC	Diethyl pyrocarbonate
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleoside triphosphate
DTT	Dithiothreitol
EDTA	Ethylene diamine tetra acetate
g	Gravity
GLB	Gel loading buffer
h	Hour
HRPO	Horseradish peroxidase
kb	Kilobase pair
kDa	Kilo dalton
mg	Milligram
min	Minute
ml	Milliliter
mM	Millimolar
nm	Nanometer
ng	Nanogram
^oC	Degree centigrade
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PHMB	p-hydroxymercuribenzoate
PMSF	Phenyl methyl sulphonyl flouride
RNA	Ribonucleic Acid
rpm	Revolutions per minute
SDS	Sodium dodecyl sulphate
s	Second
TE	Tris-EDTA

TBE	Tris-Borate EDTA
TEMED	N,N,N′,N′, Tetramethylethylenediamine
Tyr	Tyrosine
Tris	Tris (hyroxymethyl) amino ethane
U	Unit
UTR	Untranslated region
v/v	volume/volume
w/v	weight/volume
nt	Nucleotide

Contents

Introduction

Entamoeba histolytica is a microaerophilic protozoan parasite that lives in the human intestine and causes intestinal and extraintestinal amoebiasis. Some species within this genus are harmless, while others are pathogenic, causing a serious public health problem, especially in developing countries (Brumpt, 1925). It is estimated that worldwide approximately 50 million people get infected with *E. histolytica*, causing 40 thousand to 1 lakh death per year (Haque *et al.*, 2003; Huston, 2004). Amoebiasis ranks second after malaria in terms of deaths caused by parasitic protozoans worldwide (Stanley 2001, 2003; Gonzalez-Salazar *et al.*, 2009).

1.1 Classification

Kingdom: Protozoa

Phylum: Amoebozoa

Class: Lobosea

Order: Amoebida

Family: Entamoebidae

Genus*:* *Entamoeba*

Species: *histolytica*

1.2 *Entamoeba histolytica:* General features and historical perspective

E. histolytica, unlike some other protozoan parasites, has a relatively simple life cycle. It exists as either the infectious cyst form outside the human body or the invasive trophozoite in the human colon. The trophozoite divides actively by binary fission, while the cyst is dormant. Of the species of *Entamoeba* that inhabit the human intestine, *E. histolytica* is pathogenic while others like *Entamoeba dispar* is non-pathogenic. Historically, Fedor Aleksandrovich Lösch was the first who identified *E. histolytica* as a causative agent of dysentery (Losch, 1875). It was differentiated from *Entamoeba coli* and the name *Entamoeba histolytica* was assigned by Schaudinn on the basis of its ability of tissue destruction (Schaudinn, 1903). Later Brumpt suggested that there are two closely related and morphologically similar species *Entamoeba histolytica*; pathogenic and *Entamoeba dispar*; non- pathogenic (Brumpt, 1925). This view was not generally accepted for a long time. It was revisited by the isoenzyme analyses done by Sargeaunt and colleagues in which they could differentiate isolates from asymptomatic and diseased individuals (Sargeaunt and Williams, 1978). It was further confirmed by sequencing analysis of highly conserved genes and 18s rRNA that *E. histolytica* and *E. dispar* are two different species, one is pathogenic and the other is commensal, respectively (Clark and Diamond, 1991). Genome sequencing of *E. dispar* showed 90% sequence

similarity with *E. histolytica* and few species-specific genes (Lorenzi *et al*., 2010). A new species *Entamoeba nutalli* has been sequenced and phylogenetically positioned between *E. histolytica* and *E. dispar* (Tachibana *et al*., 2007, 2009). Genus *Entamoeba* includes many other important species like *E. moshkovskii, E. gingivalis, E. invadens, E. hartmanii* etc.

1.3 *E. histolytica:* morphology

The parasite has two forms; an invasive trophozoite form that occurs in the human colon and the resistant vegetative cyst form occurs outside the human body. The cyst has round shape usually having four nuclei with glycogen and chromatoid bodies. They are generally 10-15µm in diameter and encased in a refractile wall which is made up of chitin. The trophozoites are of an irregular shape and highly motile, hence their diameter varies between 10-50µm. Organelles like mitochondria, Golgi apparatus, and rough endoplasmic reticulum are typically absent. However, evidence of ER specific enzymes has been noted in the parasite (Saab *et al*., 2004; Girard-Misguich *et al*., 2008; Weber *et al*., 2008). The nucleus is spherical in shape and is 4-7µm in diameter (Clark, 2000). It is covered by a double layered membrane and shows an even distribution of Chromatin clumps. Although, the existence of mitosomes which are mitochondria-like organelles (Tovar *et al*., 1999) have been suggested to contain DNA (Ghosh *et al*., 2000). Later it was suggested that genome was not found in mitosome (Leon-Avila and Tovar, 2004).

1.4 *E. histolytica*: Life cycle and disease

The parasite acquires two stages namely cysts and trophozoites in its life cycle. Cysts are the dormant and infective form whereas, trophozoites are an invasive and motile form (Fig.1). Infection from *E. histolytica* normally occurs after ingestion of faecally contaminated food and water which contained mature cysts. The cyst travels to the highly acidic environment of the stomach through the food pipe. The quadrinucleate cyst excysts by disruption of the chitin wall after reaching to the small intestine. It goes through one round of nuclear division followed by three rounds of cellular division and gives rise to eight uninucleate trophozoites. The trophozoite migrates to the large intestine where they reside in the ileocaecal region and ingest the bacterial flora. The trophozoites divide by binary fission and re-encyst. The cysts finally pass through the faeces and completing its life cycle. The trophozoites can also come out in the stool but are not able to survive in the harsh conditions outside the human body whereas, cysts remain viable in the humid environment and can stay infective for several days. Mostly infections are non-invasive and asymptomatic; it can subside within a short period of time. The disease could be chronic non-invasive or develop into invasive disease invading the epithelial layer. This clinical syndrome is called amoebic colitis or dysentery.

Symptoms include frequent stools with mucus and occasionally blood. In severe condition, it could also spread via the bloodstream to other organs such as lungs and liver and rarely to the brain. The most common form of extraintestinal amoebiasis is termed as “abscess” in which it affects the liver and results in maximum cases of deaths due to this form of disease. Tissue invasion is not a part of the productive life cycle and therefore should be viewed as aberrant behavior of the organism.

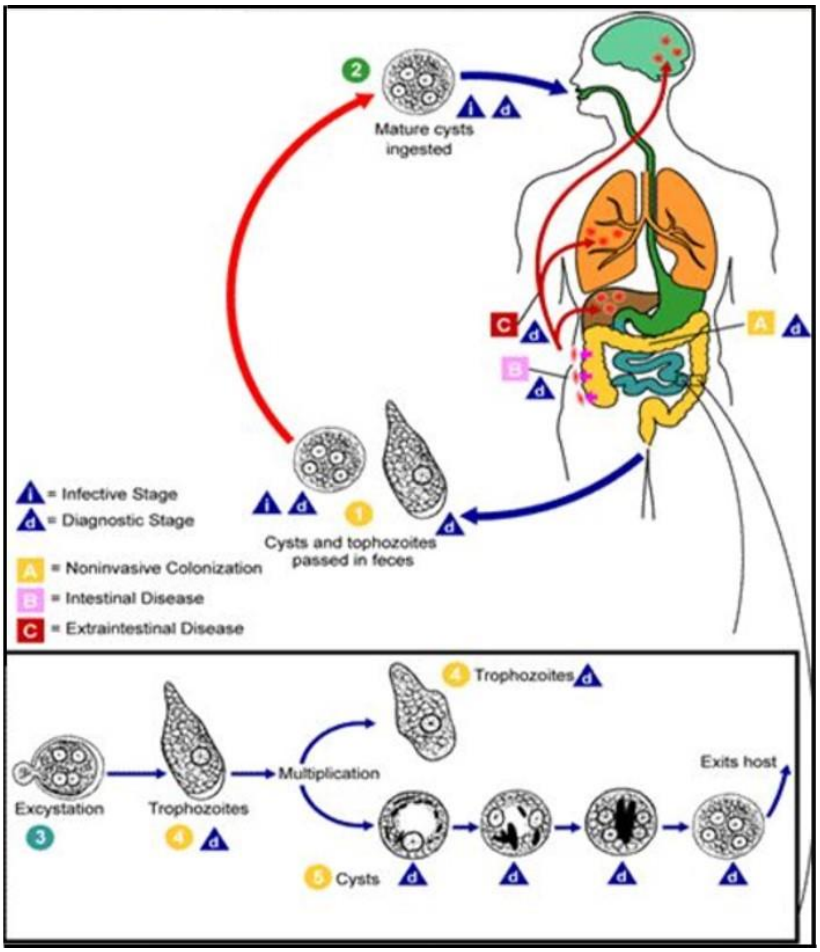


Figure 1: Life cycle of *Entamoeba histolytica*. Infection appears by ingestion of mature cysts via faecally contaminated food, water or hand. Excystation result in release of trophozoites in the small intestine which migrates to the large intestine. Trophozoites multiply by binary fission and produce cysts, and both stages pass in the faeces. Cysts can survive days to weeks in the external environment due to the protection conferred by their walls and are responsible for transmission. Trophozoites passed to the stool destroy rapidly and if ingested would not survive in the gastric environment. In various cases, trophozoites remain restricted to the intestinal lumen of individuals and cause noninvasive infection resulting in the asymptomatic carrier by passing cysts in their stool. Whereas in some patients, trophozoites invade the intestinal mucosa (intestinal disease), or through the blood stream, extraintestinal sites such as liver, brain, and lungs (extraintestinal disease), with resultant pathogenic manifestations. (Taken from http://www.dpd.cdc.gov/dpdx).

1.5 Epidemiology

People developing invasive disease due to *E. histolytica* infection are only about 10% whereas, 90% are asymptomatic (Walsh 1986). The regions with poor and unhygienic living conditions exhibit more infection. In developed countries, amoebiasis tends to be diagnosed only in travelers returning from endemic areas. It has been also diagnosed among homosexuals (Weinke *et al*. 1990, Salit *et al*. 2009). Unusual modes of transmission other than the usual oral route include oral and anal sex and contaminated enema apparatus. Relatively small number of host genes have been associated with resistance to the invasion of this parasite (Hamano *et al*. 2008) along with HLA locus and leptin genes (Duggal *et al*. 2011; Mackey-Lawrence *et al*. 2013).

1.6 Organization of nucleus in *E. histolytica* trophozoites

Unlike other eukaryotic organelles which are poorly defined in *E. histolytica,* it has a well-defined nucleus. The nuclear membrane is 120 nm thick and nuclear pores are of 50-65nm in diameter (Ludvik and Shipstone, 1970). Irregular, discontinuous clusters of "peripheral chromatin" which are composed of dense particles line the inner membrane (Miller *et al*., 1961; Ludvik and Shipstone, 1970). The nucleolus is present in the nuclear periphery (Jhingan *et al*., 2009) and is involved in the transcription and processing of rRNA genes. Although intracellular microtubules have been observed under EM, a typical mitotic spindle has not been found (Orozco *et al*., 1988). The parasite shows a unique cell division cycle which differs from other organisms (Das and Lohia, 2002; Lohia *et al*., 2007; Mukherjee *et al*. 2009). Novel proteins such as EhKlp5 (Dastidar *et al*., 2007) have been found to be involved in microtubular assembly and Formins play role in regulation of the genome content and cell division (Majumder and Lohia, 2008). Recently, a calcium binding protein EhCaBP6 has been shown to be present in microtubular ends and involved in the bridge formation during cytokinesis (Grewal *et al*., 2013).

1.7 Genome of *E. histolytica*

E. histolytica HM-1: IMSS first assembly and annotation of the genome was reported in 2005 (Loftus *et al*. 2005; Clark *et al*. 2007) that was further reassembled in 2010 and included new annotation and more sequence data. This reassembly is available on Pathema database (http://pathema.jcvi.org/Pathema/) (Lorenzi *et al*. 2010; Brinkac *et al*. 2010). The latest assembly data has been made available on Amoebadb (http://www.amoebadb.org) which is part of the EuPathDB web resource (Aurrecoechea *et al*., 2007, 2010, 2011).

The genome size of parasite is approximately 20 Mbp. Pulse-field gels predicted 14 chromosomes ranging in sizes from 0.3 to 2.2 Mbp and possibly a ploidy of 4 (Willhoeft and Tannich, 1999). The genome assembly remains fragmentary, containing 1496 scaffolds, most likely due to the high number of repetitive elements in the genome and have low GC content (24.2%) (Lorenzi *et al*., 2010). This size is consistent with data from pulse-field gels (Willhoeft *et al*. 2002) and kinetic experiments (Gelderman *et al*. 1971a; Gelderman *et al*. 1971b) which makes the *E. histolytica* genome comparable in size to that of *P. falciparum* (23 Mbp) (Gardner *et al*. 2002), *Trypanosoma brucei* (26 Mbp) (Mac leod *et al*. 2005) and the free-living amoeba *Dictyostelium discoideum* (34 Mbp) (Eichinger *et al*., 2005). The number of genes has reduced from 9985 to 8201 after the reannotation of *E. histolytica* genome largely due to the removal of apparently wrongly identified paralogues, truncated genes and very short gene models (Lorenzi *et al*. 2010) (Table 1). The ribosomal RNA (rRNA) genes show unusual organization as they are carried exclusively on 24 Kbp circular episomes (Bhattacharya *et al*. 1988) which have two transcription units as inverted repeats and are believed to account for 10-20% of cellular DNA. Some extrachromosomal DNA molecules of different sizes with no defined function have also been identified (Dhar *et al*., 1995; Lioutas and Tannich, 1995). Another unusual characteristic of the *E. histolytica* genome is the structure and organization of the tRNA genes that are organized into distinct arrays (500bp to over 1750bp), of which there are approximately 4500 in the genome (Clark *et al*. 2006, 2007).

A		
Genome	**New E. histolytica assembly**	**Old E. histolytica assembly**
Size (bp)	20799072	23361983
GC Content (%)	24.2	24.1
Number of Genes	8201	9985
Mean Gene Length (bp)	1260.9	1170.7
Number of Genes/10 Kbp	3.9	4.3
Longest Gene (bp)	15,210	15,210
Shortest Gene (bp)	147	96
Percent Coding (%)	49.7	50
Percent Genes with Introns (%)	24.4	24.9
Exons	**New E. histolytica assembly**	**Old E. histolytica assembly**
Number	10,754	13,176
Mean number per Gene	1.3	1.3
GC Content (%)	28	28.1
Mean Length (bp)	962	886.1
Total Length (bp)	10,340,284	11,675,669
Introns	**New E. histolytica assembly**	**Old E. histolytica assembly**
Number	2,553	3191
GC Content (%)	19.3	21.7
Mean Length (bp)	74.1	100
Total Length (bp)	189,260	319,223
Intergenic Regions	**New E. histolytica assembly**	**Old E. histolytica assembly**
GC Content (%)	20.5	20
Mean Length (bp)	708.7	823.5

B			
Annotation	**OGA**	**NGA**	**NGA-curated genes***
Genes with EC number	124	1098	604
Genes with GO terms	3106	3468	1843
Number of domains	816	1347	317

Table 1: *Entamoeba histolytica* genome sequence analysis (A) Comparative genome statistics between old and current assembly (B) Comparative view of EC number, GO term and domain identification between old and new annotations. OGA: original genome annotation; NGA: new genome annotation. (Adopted from Lorenzi *et al.*, 2010).

Initially homologous recombination machinery was reported to be absent in *E. histolytica* but reports suggest the presence of this machinery. Some of the key proteins, such as EhRad51, EhRad54, and EhBLM are reported to have a role in DNA repair (López-Casamichana *et al*. 2008; Charcas-Lopez Mdel *et al*. 2014). Experimentally, homologous recombination has been demonstrated in *E. histolytica* which was stimulated in response to growth stress (Singh *et al*. 2013). The presence of meiosis-specific genes indicates the possibility of meiosis.

1.7.1 Gene organization

E. histolytica is expected to exhibit high gene density due to small genome size. About 25% of genes show the presence of introns (mostly single) and 6% contain multiple introns (Clark *et al.*, 2007; Loftus *et al*., 2005). Small sets of linked genes (gEH-FeSOD, gEH-AP and gEH-170; RP-

L21 and actin) have been studied suggesting tight packing as their intergenic regions were found to be between 0.4-2.3kb (Bruchhaus *et al*., 1993; Petter *et al*., 1992). The introns identified in the parasite are relatively short and are AT rich compared to the corresponding exon sequences. In contrast to higher eukaryotes, *Entamoeba* introns do not contain a well-conserved branch point consensus and have extended donor and acceptor splice sites sequences (Wilihoeft *et al*., 2001). In higher eukaryotes, alternative splicing and polyadenylation, are the major mechanisms for expanding the diversity of their transcriptomes and proteomes (Keren *et al*., 2010; Ozsolak *et al*., 2010). In human, 95% of multi-exon genes undergo alternative splicing (Pan *et al*., 2008; Wang *et al*., 2008) and at least 42% of intron-containing genes are alternatively spliced in *Arabidopsis thaliana* (Filichkin *et al*., 2010). Moreover, microheterogeneity (Pauws *et al*., 2001) of polyadenylation site usage in eukaryotic mRNAs are also found to be extensive. Although numerous studies demonstrated the pervasiveness of alternative splicing and polyadenylation in higher eukaryotes, transcriptome analysis determined that the functional relevance of these events is limited to a small proportion in *E. histolytica* genes (Hon *et al*., 2013). In *Entamoeba*, the untranslated 5' and 3'-regions of structural genes are usually short, consisting of 5-21bp and 14-44bp respectively (Bruchhaus *et al*., 1993). Few genes have been reported to have long untranslated regions in the 5'and 3'-ends (De Meester *et al*., 1991; Gangopadhyay *et al*., 1997b; Urban *et al*., 1996). Comparison of the 5' flanking regions from 37 protein coding genes of *E. histolytica* indicated the presence of three conserved motifs in the core promoter region (Purdy *et al*., 1996). One of the three conserved sequences, (GTATTTAAAG/C) lies at -30 from the transcription start; the second, (AAAAATTCA), overlies the transcription start; and the third element, (GAAC), lies at variable location between the first two sequences. The region at -30 (TATA box) was shown by positional analysis to control the site of transcription initiation and has been shown to be bound by a factor in *E. histolytica* nuclear extracts (Bruchhaus *et al*., 1993). The region overlying the transcription initiation site (*Inr*) also controls the site of transcription initiation but appears to play a lesser role than the other two sequences (Singh *et al*., 1997). The presence of unusual TATA and *Inr* elements and the presence of a novel third element indicate that the basal transcription machinery of *E. histolytica* differs significantly from that of other eukaryotes. This is not surprising considering that mammalian viral promoters do not function in *E. histolytica* and amebic promoters do not function in mammalian cells (Purdy *et al*., 1994). A putative TATA binding protein (TBP) has been reported for *E. histolytica* (GenBank Acc. No. Z48307) which is significantly divergent from the TBP of higher eukaryotes (Hernandez *et al*., 1997), as also from that of the protozoan parasite, *Plasmodium falciparum* (McAndrew *et al*, 1993). Transcription of protein-encoding genes in *E. histolytica* is resistant to α-amanitin (Lioutas and Tannich, 1995), in contrast to RNA

polymerase II from higher eukaryotes. This may reflect the variant structure of the *E. histolytica* polymerase II subunit, analogous to that of another early divergent protozoan *Trichomonas vaginalis* with α-amanitin resistant RNA polymerase II (Quon *et al*., 1996).

1.7.2 Repetitive DNA elements

The genomes of protozoan parasites contain repetitive DNA in varying extent (Bhattacharya *et al*., 2002; Wickstead *et al*., 2003). In *E. histolytica genome,* various transposable element (TEs) families have been described. Among all, thoroughly studied repeat family is non-Long Terminal Repeat (LTR) retrotransposons. Non-LTR retrotransposons contain three subfamilies of LINE (long interspersed nuclear element) (EhLINE1, EhLINE2 and EhLINE3), and SINE (short interspersed nuclear element) (EhSINE1, EhSINE2 and EhSINE3). Moreover, a *mutator*-related DNA transposon and one novel element ERE1 have been reported in *E. histolytica* and *E. dispar* (Bakre *et al*., 2005; Lorenzi *et al*., 2008; Pritham *et al*., 2005; Sharma *et al*., 2001, Van Dellen *et al*., 2002). Another element ERE2 was also reported in *E. histolytica* (Lorenzi *et al*., 2008). In *Entamoeba* species, the ribosomal RNA genes are present on rDNA plasmids, which constitute the major repetitive DNA (about 10% of the genome) (Bhattacharya *et al*., 1988). In *E. histolytica* the rDNA plasmid (EhR1) contains families of short repetitive DNA elements, such as the *ScaI (*Mittal *et al.,* 1992*)* and *DraI* (Mittal *et al*., 1991) repeats. *DraI* repeats are highly recombinogenic and exhibit length variation within an isolate (Mittal *et al*., 1994). Several other repeat sequence families have also been reported in *E. histolytica* (Bhattacharya *et al*., 1988; Lohia *et al*., 1990; Michel *et al*., 1992; Mittal *et al*., 1994). Like other parasites, some protein coding genes in *E. histolytica* contain internal repeated structures which show intraspecific repeat variation. For example, in the case of chitinase, repetitive sequences are found near the amino terminus of the protein that varies in both length and sequence among isolates (de la Vega *et al*., 1997). Other repetitive elements are also present in the genome whose functions are yet to be elucidated. There are over 75 genes encoding leucine-rich tandem repeats (LRR) of the type found in BspA-like proteins of the *Treponema pallidum* LRR (TpLRR) subfamily, which has a consensus sequence of LxxIxIxxVxxIgxxAFxxCxx (Davis *et al*., 2006) along with stress sensitive protein (Ehssp) (Satish *et al*., 2003).

1.8 Transposable elements

Transposable elements (TEs), also known as mobile genetic elements, were first discovered by geneticist Barbara McClintock in 1950s (Craig *et al*., 2002); since then various types of TEs have been discovered in most of the prokaryotic and eukaryotic genomes. These are DNA sequences

which have the ability to move within a genome to non-homologous insertion sites (Craig, 1997). Thus, they re-shape the genome by rearrangement, with the potential to shuffle existing genes and modulate their expression pattern. TEs are a ubiquitous, abundant and diverse component of eukaryotic genomes that occupy 50% or more of the genome content in some organisms; up to 80% of nuclear DNA in plants, 3 to 52% in metazoans and 3 to 20% in fungi (Wicker and Keller, 2007). They duplicate themselves in the process of insertion into new chromosomal locations and result in a variation of chromosome size. It is well known that TEs have a crucial function in nuclear architecture, stability of the genome, amplification of genes, and altered regulation of genes (Deininger *et al*., 2003; Slotkin *et al*., 2005). They are able to influence or alter the genome in many ways and can introduce a variety of changes in gene structure and expression (Thomas *et al*., 2010). Some of the ways TEs can impact the genome are shown in figure 2.

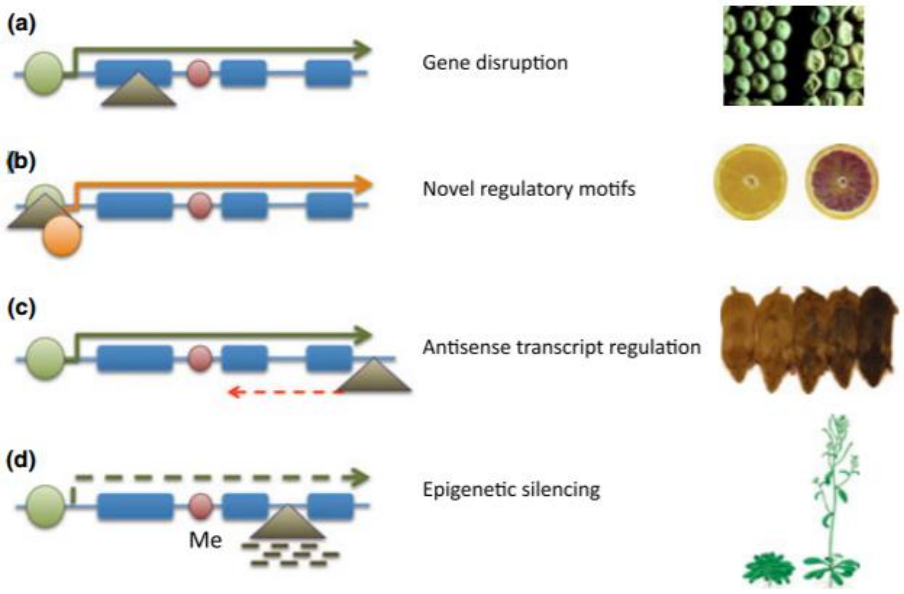


Figure 2: Impact of TEs insertions on gene structure and function. (a) Insertion of TE into the coding regions abolished gene function and resulted in wrinkled-seed pea as described by Mendel. (b) TEs inserted into the regulatory region of a gene can work as a promoter and altering normal expression of the gene, which is responsible for the blood oranges as shown. (c) Gene transcription levels can be inhibited by antisense transcription from adjacent TE insertions as observed for the agouti colour gene in mice (d) Insertion of a DNA transposon into the first intron of FLC (a gene that delays flowering) targeted by TE-derived siRNAs results in gene silencing and early flowering in *Arabidopsis thaliana*. Exons are shown as blue rectangles; TE insertions as triangles; gene promotors and TEs are shown as green and orange circles respectively. Arrowheads indicate gene transcription (Adopted from Bonchev *et al*., 2013).

1.9 Classification of Transposable elements

TE classification system was first proposed by Finnegan in 1989. They are classified into two classes, Class I and Class II on the basis of their transposition intermediates. Class I transposons are also called retrotransposons that transpose by "copy and paste" mechanism. They transpose by an RNA intermediate which reverse transcribes into cDNA and inserts within the host genome. Retrotransposons are subdivided into two major groups on the basis of long terminal repeats (LTRs). The first group contains the LTR and is represented by LTR retrotransposons; tyrosine recombinase retrotransposons, and endogenous retroviruses. The second group contains non-LTR retrotransposons; long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and processed pseudogenes being the main representatives of this group. A third group which constitutes a novel class of eukaryotic retroelements, are known as *Penelope-like* elements and are distinct from both non-LTR and LTR retrotransposons. They have been first identified in *Drosophila virilis* (Evgen'ev *et al*., 1997) and later in other eukaryotes. Class II transposons are DNA transposons that do not involve RNA intermediate and mobilize themselves by moving their DNA sequence using 'cut and paste' mechanism.

1.9.1 LTR retrotransposons

LTR retrotransposons contain diverse elements with long terminal repeats in their sequence. LTR-retrotransposons show similarity with the retroviruses on the basis of their structure except for the absence of *env* (envelope) gene in most elements (Eickbush and Jamburuthugoda, 2008). They contain *gag* gene which encodes a structural protein and has nucleic acid binding activity; *pol* gene, which is a polyprotein encoding protease (PR), reverse transcriptase (RT), RNaseH (RH) and integrase (IN) activities (Gogvadze and Buzdin, 2009). These have been divided into three major classes on the basis of the phylogeny of their RT domain; the Ty1/copia family, Ty3/gypsy and BEL families (Fig.3A). The Ty1/copia and Ty3/gypsy elements are extensively distributed throughout plants, animals, and fungi, whereas the Bel elements are reported only in animals. Other than these there are two groups of LTR containing elements known as endogenous retroviruses (ERVs) and tyrosine recombinase or YR encoding retrotransposons (Fig.3A). The YR retrotransposons encode a tyrosine recombinase rather than an integrase. They include DIRS-like elements which are flanked by inverted repeats and contain an internal complementary region, and PAT-like elements (Poulter and Goodwin, 2005). The ERVs constitutes around 1% of the human DNA and have been found in all vertebrate genomes. They encode *env* gene and are believed to have traces of ancient germ-cell retroviral infection (Sverdlov, 2000). The Copia was the first element studied in *Drosophila melanogaster* (Mount and Rubin, 1985) and Ty1 in *Saccharomyces*

cerevisiae. Utilization of RNA intermediate for the insertion of a DNA copy was first reported in yeast for the Ty1 element (Boeke *et al*., 1985). The mechanism of transposition used by LTR retrotransposons has been reviewed extensively (Boeke and Corces, 1989). Prevalence of these elements is highly variable in animals, low in fungi and high in plants as it has been reported that in maize genome 75% increase in size is a result of the proliferation of 11 families of these elements (SanMiguel *et al*., 1998).

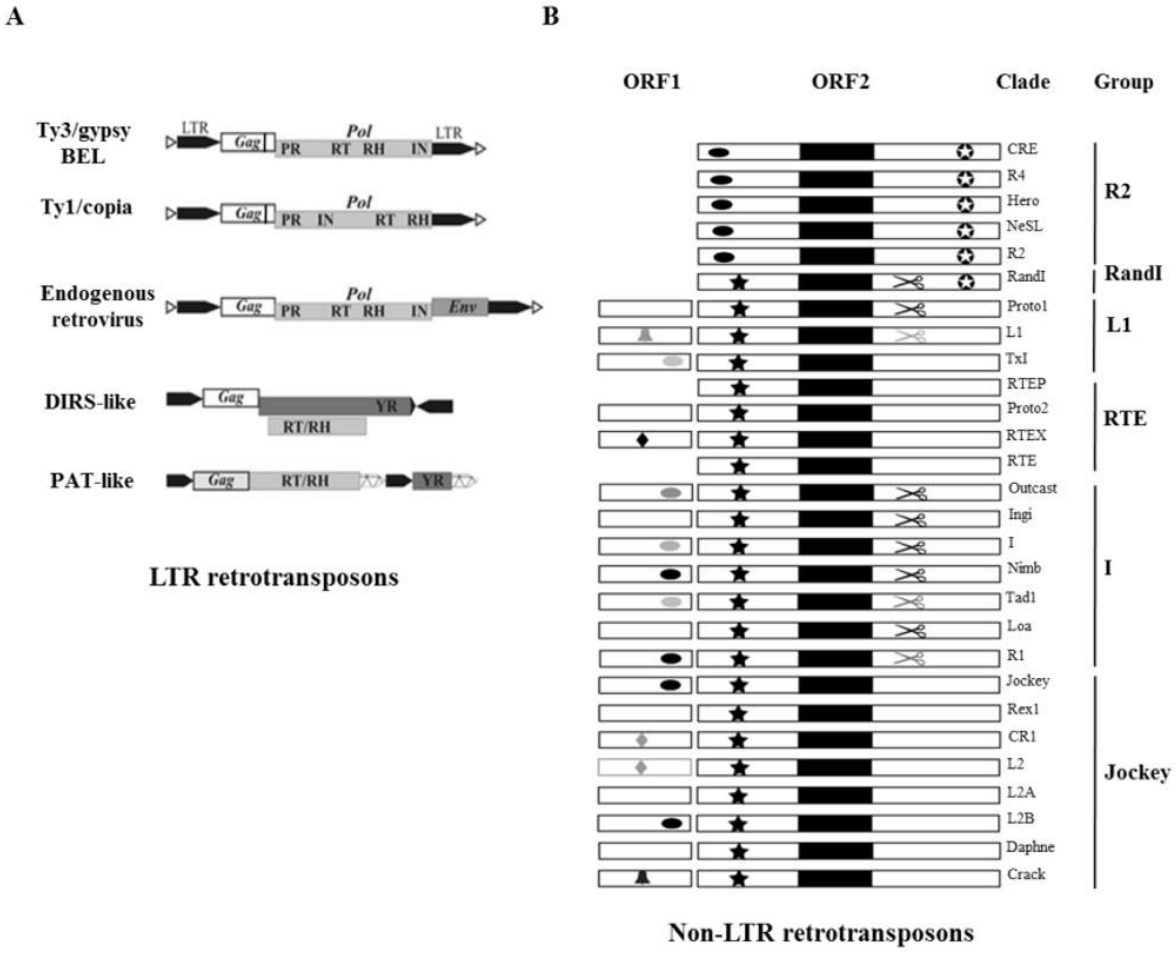


Figure 3: (A) Schematic representation of various domains in LTR retrotransposons. The domains and other structural components of elements of some major groups are represented as follows- white triangles; short direct repeats (TSD), black pentagon; LTR, PR; protease, RT; reverse motifs which are present only in some families. (Adapted from Gogvadze and Buzdin, 2009). (B) Schematic representation of ORF1 and ORF2. The ORF1 and ORF2-encoded proteins are shown as short and long white rectangles. In the ORF2 proteins, black rectangles mark the RT domains; black and white asterisks denote the AP and REL-ENDO type endonucleases, respectively; scissors denote RNase H. In the ORF1 proteins, bells and diamonds mark the esterase (Kapitonov and Jurka, 2003) and RRM domains respectively (Kapitonov and Jurka, 2005; Khazina and Weichenrieder, 2009). In ORF1 and ORF2 proteins, ovals represent cysteine-histidine motifs. Domains and ORF1 that are present only in some families of a particular clade are shown in gray. RH; RNaseH, IN; integrase, *Env*; envelope, YR; tyrosine recombinase. Black vertical bar represents cysteine-histidine (Adapted from Kapitonov *et al*., 2009).

1.9.2 Non-LTR retrotransposons

As the name indicates, non-LTR retrotransposons are devoid of long terminal repeats and many of them end with a poly(A) tail at their 3'-ends, whereas their 5'-ends often contain variable deletions (5' truncations). These retrotransposons are highly abundant in eukaryotes. They are also known as poly(A) retrotransposons, or target-primed retrotransposons (Han, 2010). Non-LTR retrotransposons have been divided into 6 groups namely, R2, RandI, L1, RTE, I and Jockey. These groups are further divided into different clades (Fig.3B). The highly abundant TEs of this group analyzed in mammals are LINEs (long interspersed nuclear elements), to differentiate them from SINEs (short interspersed nuclear elements) (Eickbush *et al.* 2008). As all LINEs possess the RT and other activities, are proposed to be autonomous and retrotranspose via their own machinery.

The non-LTR retrotransposons consist of centrally localized RT domain and C-terminus REL-type endonuclease (EN) domain have been placed into R2 group which has been further classified in various clades like R2, R4, NeSL, CRE, Hero etc (Fig.3B). The RT domain of *E. histolytica* element EhLINE1 exhibits similarity with the RT domain of R4 clade elements (R4 element of *Ascaris lumbricoides*, Dong element of *Bombyx mori*) and its EN domain shows similarity with the EN domain of elements of R2, R4, and CRE clades. Among non-LTR retrotransposons, some groups encode a single ORF, while others contain two ORFs. In elements which encode a single ORF, the N-terminal region of the polypeptide displays nucleic acid binding activity, while in elements with two ORFs, the ORF1 contains nucleic acid binding activity and ORF2 contains RT and EN activities. For example, an N-terminal region of the R2 element of *B. mori* and Horseshoe crab (Christensen *et al*., 2005), a NeSL element of *C. elegans* (Malik and Eickbush, 2000) bear nucleic acid binding domain (CCHC and C-myb) and all of which code for a single polypeptide. The CZAR and SLACS element of *T. cruzi* and *T. brucei* respectively contain two ORFs and their first ORF displays CCHC nucleic acid binding motifs (Aksoy *et al*, 1990, Gabriel *et al*, 1990, Villanueva *et al*, 1991). The R4 element of *A. lumbricoides* and Dong element of *B. mori* also contain a long N-terminus region upstream of the RT domain, but it is not known whether they have nucleic acid binding activities.

1.10 Autonomous retrotransposon: Long Interspersed Nuclear Elements (LINEs)

As mentioned above, LINEs are autonomous non-LTR retrotransposons. These are widespread in eukaryotic genomes and the major LINE, L1 constitutes around 17% of the human DNA (Lander *et al*., 2001). Of the 500,000 L1s in the human genome, around 7,000 are full-length and of those only 80-100 are predicted to be active for retrotransposition (Brouha *et al*., 2003). Similar is the case

with LINEs of other organisms as well. Human L1 is the most extensively studied LINE so far. The full-length L1 is 6 kb long; has 900nt 5'-UTR with an internal promoter driven by RNA polymerase II, two open reading frames (ORF1 and ORF2), a short 3'-UTR and poly(A) tail. An antisense promoter has also been mapped in the 5'-UTR (Erwin *et al*., 2014) (Fig.4).

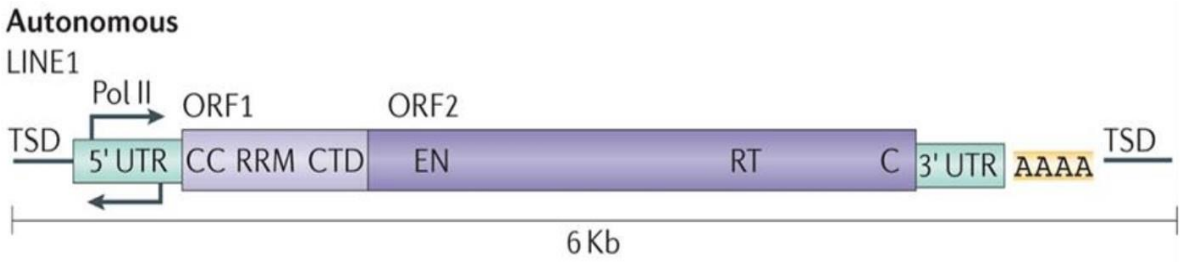


Figure 4: Autonomous element (LINE1) in human. In humans, LINE1 is of 6kb long with a strong promoter located at its 5'-UTR along with a weak antisense promoter. L1s consist of two ORFs; ORF1 and ORF2. ORF1 encodes a 40kDa RNA binding protein containing coiled coil (CC), RRM and a C-terminal domain (CTD) and ORF2 encode a 150kDa protein with endonuclease (EN), reverse transcriptase (RT) and a cysteine rich (C) domain. L1 RNAs usually terminates via a canonical poly(A) signal (AATAAA) at the 3'-UTR can also frequently bypass this termination signal for a downstream poly(A) signal in the 3'-flanking DNA. L1s genomic insertion terminate in a varying length of the poly(A) tail (AAAn) and are flanked by a TSD (4-16bp in length, black horizontal arrow) (Adopted from Erwin *et al*., 2014).

All non-LTR retrotransposons possess endonuclease (EN) and reverse transcriptase (RT) enzymatic activities that are required for the transposition reaction. Both the activities are encoded by ORF2. The RT domain is maximally conserved in all elements (Malik *et al*., 1999). The EN domain is of two types: apurinic/apyrimidinic (AP) endonuclease and type IIS restriction enzyme-like endonuclease (REL-ENDO). All non-LTR retrotransposons reported either have AP endonuclease or REL endonuclease. Exceptionally in *Chlamydomonas reinhardtii*, RandI/Dualen element encodes both REL and AP type endonucleases (Fig.3B) (Kojima and Fujiwara, 2005). Details about the LINE encoded proteins are given below:

1.10.1 ORF1p

The first ORF (ORF1) in non-LTR retrotransposons with two ORFs encodes a protein which can bind to nucleic acids (Martin, 1991). Some elements such as “I and Jockey” contain ORF with three cysteine-histidine motifs (CCHC type) that are considered to be associated with the nucleic acid binding domain and exhibit similarity with the “gag” protein of many LTR retrotransposons. Elements like R2 have single ORF whose N-terminal region contains conserved C_2H_2 zinc-finger

and/or c-myb DNA binding motifs (Yang *et al*., 1999) which are equivalent to ORF1 of the elements with two ORFs. Such CCHC motifs are not present in the human L1 element instead, it contains a leucine zipper domain which is located centrally and required for the multimer formation of ORF1p (Hohjoh and Singer, 1996). The C- terminus of human L1 ORF1p contains highly basic conserved amino acid residues which are thought to function in RNA binding. A non-canonical RNA-recognition motif (RRM) that is the most common eukaryotic RNA-binding domain has been identified in ORF1p of mammalian L1 and in some *gag*-like ORF1 proteins (Khazina and Weichenrieder, 2009). The role of ORF1p in retrotransposition has been confirmed by missense mutations in either the leucine zipper or conserved carboxyl terminal amino acids, which abolish L1 retrotransposition in cultured cells (Moran *et al*., 1996). L1 ORF1p colocalizes with the full-length sense strand L1 RNA in cytoplasmic ribonucleoprotein particles (RNPs), which are proposed to be the intermediates in L1 retrotransposition as revealed by biochemical studies in teratocarcinoma cells (Hohjoh and Singer, 1996). In addition, ORF1p has been shown to co fractionate with full-length L1 RNA in mice (Martin, 1991) and is regulated developmentally in both spermatogenesis and oogenesis (Trelogan and Martin, 1995). Recombinant ORF1p has the ability to bind with both RNA and single stranded DNA in a cooperative manner *in vitro* and does not show sequence specificity (Kolosha and Martin, 1997). On the basis of comparison of ORF1 from various non-LTR elements, it may be stated that the possible function of ORF1p is to associate with the RNA template and import the template back into the nucleus for reverse transcription.

1.10.2 ORF2p

In non-LTR retrotransposons with two ORFs, it is the ORF2p which contains a reverse transcriptase (RT) and endonuclease (EN) domain that is necessary for the process of retrotransposition. The elements with one ORF also contain an RT domain, which is at the centre (Fig.3B). ORF2 encodes a 150kDa multidomain protein in human and contains a centrally localized RT and an EN domain. Other than RT and EN domain, there are two regions in the ORF2p shown to be important for retrotransposition known as Z domain and Cysteine rich (Cys) domain (Fig.5). The Z domain is located adjacent to the EN domain and contains a conserved octapeptide sequence (Clements *et al*., 1998). In related R2 elements, this octapeptide sequence has been shown to be a part of an RNA binding motif (Jamburuthugoda *et al*., 2014). In addition, Z domain also contains a PCNA binding motif which has been shown to be important for retrotransposition (Taylor *et al*., 2013). ORF2p also possess a C-terminal Cys domain shown to be important for L1 retrotransposition and may have a role in nucleic acid binding as mutation in Cys domain can cause low rate of *in vitro* retrotransposition and RNP formation (Piskareva *et al*., 2013; Moran *et al*., 1996).

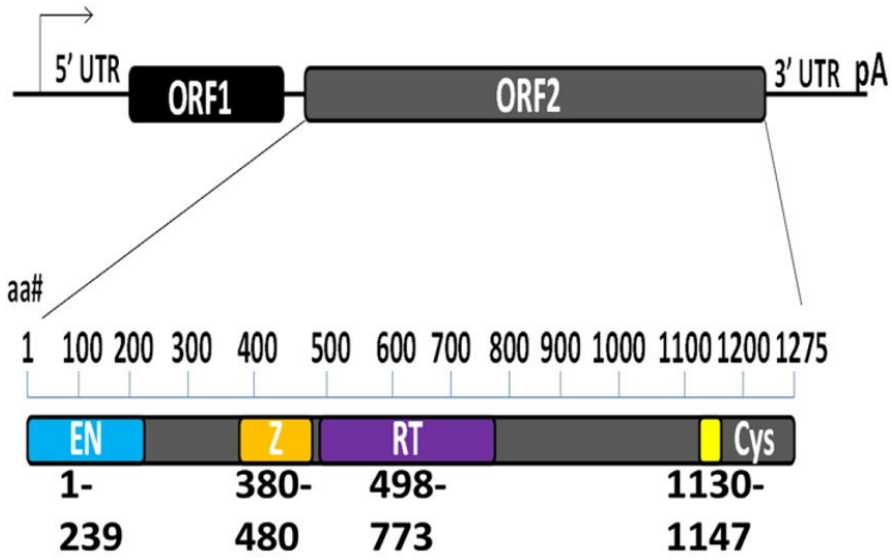


Figure 5: Long Interspersed Element 1 (LINE-1/L1) and ORF2 in the human genome. A schematic representation of L1, including the transcription start site in the 5'-UTR, ORF1, ORF2, the 3'-UTR and poly(A) signal. The inset shows the schematic of ORF2p (drawn to scale with amino acid number). The ORF2p contain an EN domain (blue), Z domain (orange) including a PCNA binding sequence, RT domain (purple) and Cys domain (yellow) that may be involved in nucleic acid binding. These domains exhibit ~50% of the ORF2p sequence and have both catalytic (EN and RT) and noncatalytic functions (Z and Cys) important for retrotransposition. The rest of the ORF2p (~50%: gray) has no known function in retrotransposition. Amino acid positions are mentioned below each domain (Adopted from Christian *et al.*, 2016).

1.10.2.1 The Reverse transcriptase domain

It has been demonstrated by Phylogenetic analysis that RTs encoded by non-LTR retrotransposons are distinct from the RTs encoded by LTR retrotransposons and retroviruses with the former being more primitive (Malik *et al.*, 1999). Non-LTR retrotransposon-encoded RT displays both RNA and DNA-dependent polymerase activities (Garcia-Perez *et al*, 2003). The DNA polymerase activity allows second strand DNA synthesis after reverse transcription. RT can incorporate dNTPs in a poly(dA)/oligo(dT) or poly(dC)/oligo(dG) template/primer set. A highly conserved YXDD motif is present in RT domain of both LTR and non-LTR retrotransposons. It is required for RT activity as a mutation in D residues has been shown to abolish the RT activity (Mathias *et al.*, 1991; Moran *et al.*, 1996; Yang *et al.*, 1999). The RTs encoded by non-LTR retrotransposons in some of the organisms have been summarized below:

(i) *Bombyx mori* R2 RT

RT activity of the ORF of R2Bm expressed in *E. coli* has been studied (Luan *et al.*, 1993). This 120 kDa protein could bind to both RNA and DNA. In non-LTR elements, retrotransposition takes place via target-primed reverse transcription (TPRT) (Fig.7B) (Christensen *et al.*, 2006) in which RT uses the 3'-OH of DNA generated by the bottom strand cleavage of the target site to prime reverse

transcription of R2 RNA. Target site DNA cleavage is carried out by EN activity of the R2Bm. R2 RT can add non-template nucleotides (usually T residues) if there are deletions at the 3'-end (Luan and Eickbush, 1995). It has high processivity on both RNA and single stranded DNA that may enable this enzyme to synthesize both DNA strands during the process of retrotransposition. R2 RT does not have RNase H domain (Malik *et al*., 1999), and no RNase H activity has been detected in *vitro* (Kurzynska-Kokorniak *et al*., 2007). Further, as it can use either RNA or DNA strand as a template, it can efficiently displace an RNA or DNA strand. R2 RT also shows the property of template jumping and can synthesize a continuous cDNA strand on a non-continuous RNA template. It has been shown that RT resumes reverse transcription from 5'-end of one RNA template to the 3'-end of another RNA template (Bibillo and Eickbush, 2002). Studies have shown that R2 RT is a low fidelity enzyme (Eickbush and Jamburuthugoda, 2008).

(ii) Human L1 RT

Mathias *et al*., 1991 first demonstrated reverse transcriptase activity in L1 RT. It has both DNA and RNA directed polymerase activity as demonstrated by primer extension studies. Basic properties of L1 RT are similar to the R2 RT; it also reverse transcribes via nick priming. The provided nick could be generated by either ORF2 EN or by DNA lesions in the genome. Further, it can also add non-templated nucleotides during reverse transcription. The RNP particles containing L1RT have been isolated by Kulpa and Moran in cultured mammalian cells (Kulpa and Moran, 2006).

(iii) *Trypanosoma cruzi* L1

Trypanosoma L1Tc encodes a single ORF and displays both RNA- and DNA-directed reverse transcriptase activities (Garcia-Perez et al., 2003) and template jumping property like R2 RT.

1.10.2.2 The Endonuclease domain

Other than the RT, ORF2p also has EN domain which shows endonuclease activity and is required for retrotransposition. On the basis of EN domain, non-LTR retrotransposons can be divided into two subtypes (Yang *et al*., 1999). First one is characterized by having two ORFs and an EN domain that closely resembles the apurinic/apyrimidinic (AP) endonuclease which is involved in DNA repair (Feng *et al*., 1996; Martin *et al*., 1995). Elements of the second type have single ORF with a restriction enzyme-like endonuclease (REL-ENDO). There are some exceptions in which a single ORF element may have AP type EN domain, and vice versa. Also, some elements such as RandI/Dualen elements of *C. reinhardtii* contain both types of EN domains. AP endonuclease (APE) is the more common in non-LTR retrotransposons; whereas REL-ENDO is only present in R2 and RandI groups (Fig.3B). APE functions in the repair of apurinic/apyrimidinic sites, 3'-

phosphatase and 3'-phosphodiesterase activities (Demple and Harrison, 1994). *T. cruzi* (L1Tc) demonstrated all the three activities of APE (Olivares *et al*., 1999). Most members of the APE group are not site-specific whereas many are known to insert at defined target sites (DRE, Tdd-3, and related elements in *Dictyostelium discoideum*, Zepp elements of *Chlorella* spp., R1 of *Bombyx mori*) (Eickbush and Malik, 2002). Although these elements inserted at defined sites, the endonucleases encoded by them were not strictly site specific. The human L1 retroelement inserts in human DNA at many sites with a preferred nicking site for its endonuclease at 5'-TTTT↑A-3' (where the arrow denotes the nicking site). Besides the preferred sequence, L1 EN can nick other sequences as well (Cost and Boeke, 1998). In Human, endonuclease has been observed to introduce nicks in both supercoiled plasmids and oligonucleotide substrates (Feng *et al*., 1996). Sequence non-specificity is not the case with every APE containing element, some are strictly site specific. These elements insert into tRNA genes, rRNA gene clusters, snRNA genes, other transposable elements, microsatellites, and telomeric repeats. For example, in *B. mori*, the R1 element inserts into a specific location in 28S rRNA genes (Feng *et al*., 1998; Jakubczak *et al*., 1991). Non-LTR retrotransposons in *Dictyostelium discoideum* belong to the TRE family which are grouped in L1 clade and reside near tRNA genes in clusters. TRE5-A, TRE5-B and TRE5-C elements insert 40-54 bp upstream to tRNA genes (Beck *et al*., 2002; Winckler *et al*., 2002)., TRAS1 and SART1 (included in R1 clade) are the telomere-specific retroelements in *B. mori*. They insert in 5'-TTAGG-3'-telomere repeat arrays (Okazaki *et al*., 1995; Takahashi *et al*., 1997) by generating specific nicks on both strands of the telomeric repeat sequence between T and A of the (TT↑AGG)n bottom strand and C and T of the (CC↓TAA)n on top strand (Anzai *et al*., 2001). Unlike other retrotransposon-encoded EN, TRAS1 EN nicks the top strand site-specifically *in vitro*.
The second subtype of EN domain encoded by non-LTR retrotransposons is the restriction enzyme like endonuclease (REL-ENDO), which resides downstream to the RT domain. REL-ENDO domain is present in the R2 group and most of the elements in this group are site specific. R2 and R4 clade non-LTR retrotransposons insert in the 28S gene of arthropods and nematodes (Burke *et al*., 1995; Yang *et al*., 1999). Members of CRE clade, like CRE1, CRE2, SLACS and CZAR of Trypanosomes, and members of NeSL clade of Nematodes insert into specific spliced leader sequences (Aksoy *et al*., 1990; Gabriel *et al*., 1990; Malik and Eickbush, 2000). RandI elements, containing both AP and REL-ENDO type endonucleases, were first identified in *Chlamydomonas reinhardtii* and later in unassembled HTGS (High- throughput genomic sequence) of *Arabidopsis thaliana*. The single ORF present in the elements encodes RT, RNase H, cysteine protease and two endonucleases (Fig.3B).

1.10.2.3 The RNase H domain

Most non-LTR retroelements do not code for their own RNase H and utilize host-encoded RNase H for the removal of RNA from RNA-DNA hybrid during the process of retrotransposition. However, several lineages of retroelements including all LTR and some non-LTR code for their own RNase H domain (Fig.3A,B). These elements include *I* element of *D. melanogaster*, RandI/Dualen element of *C. reinhardtii* (Fawcett *et al*., 1986; Kojima and Fujiwara, 2005) and the L1 element of *T. cruzi* (Olivares *et al*., 2002).

1.11 Non-autonomous retrotransposons

1.11.1 Short Interspersed Nuclear Elements (SINEs)

Eukaryotic genomes are also invaded by short 100-500bp non-coding sequences known as SINEs. Unlike LINE elements, SINEs are non-autonomous and require a partner LINE for their retrotransposition. In spite of this they are highly successful genomic parasites, e.g. the human SINE, Alu, is present in ~ 1.5 million copies. Unlike LINEs, SINEs are derived from Pol III transcripts and lack any ORF, however they contain internal promoter similar to LINEs. Some SINEs have been shown to be derived by tRNA promoter (Ohshima and Okada, 1994) whereas some rodent and primate SINEs (rodent B1 SINEs and human Alu) are derived from a 7SL gene promoter (Fig.6) (Ullu and Tschudi, 1984).

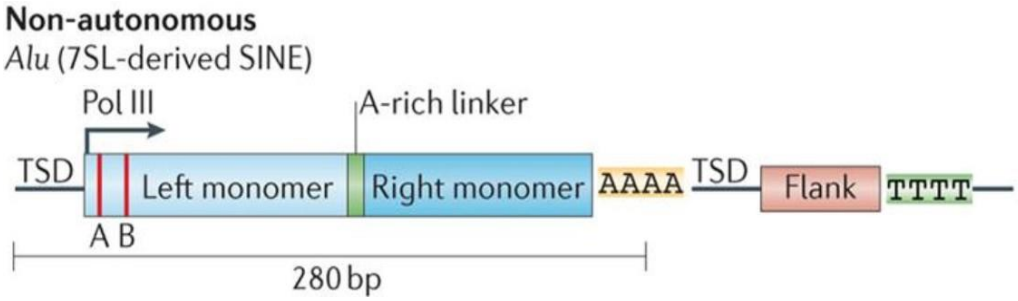


Figure 6: Non-autonomous element in Human. A full length (280 bp) Alu element, containing an internal pol III promoter (A and B box in red) at its 5'-end is shown. Alus are generated by the dimerization of two 7SL RNA sequence (left and right monomer). Alu genomic insertions terminate in a poly (A) tail and flanked by a TSD (black horizontal arrow) similar to L1. Alu transcripts terminate at pol III terminator sequences (TTTT) in the downstream flanking sequence. (Adopted from Erwin *et al*., 2014).

Many SINE elements exhibit sequence similarity at the 3'-end with their LINE partner, for example in humans, the Alu and L1 elements both terminate in poly(A) tail and are flanked by target site

duplication (TSD). Similarly, the CR1 LINE and Pol III/SINE of tortoise (Ohshima *et al*., 1996); the eel UnaL2 LINE and UnaSINE1 (Kajikawa and Okada, 2002); mammalian LINE2 and MIR SINE (Smit and Riggs, 1995) and the ruminant Bov-B LINE and Bov-tA SINE (Okada and Hamada, 1997) all display striking similarities at their 3′ends. The similarity at 3'-ends leads to the suggestion that SINEs, being non-autonomous possibly mimic the active LINE RNA, and allows the recognition by LINE retrotransposition machinery (Kajikawa and Okada, 2002; Ohshima *et al*., 1996). Experiments performed in HeLa cells with eel SINE also support this proposal. The incorporation of 3'-sequence shared by eel UnaL2 LINE/UnaSINE1 pair enhanced retrotransposition rate of a cloned element in cultured cells (Kajikawa and Okada, 2002).

1.11.2 Processed Pseudogenes

Processed pseudogenes also show 3'- sequence similarity with LINE sequences as they end in a poly(A) tail that is similar to L1 in humans and are flanked by target-site duplication (Vanin, 1985). Due to these features, it was proposed that like SINEs, pseudogenes may also be mobilized using the LINE machinery. Processed pseudogenes do not contain introns and have lost untranscribed part of their promoters (Esnault *et al*., 2000; Ohshima *et al*., 2003; Weiner *et al*., 1986). They are integrated copies of cDNAs of various cellular mRNAs. Mostly, these are non-functional or inactive and do not code for proteins due to acquired mutations in the absence of selection pressure. Sometimes, due to the presence of an active promoter upstream of the insertion site, and conservation of the complete ORF, they become active. *PGK2* and *PDHA2* are expressed and active pseudogenes in humans.

1.11.3 Penelope like elements

The elements were first discovered in *Drosophila virilis* and further found to be present in different eukaryotic genomes. They possess internal promoter and exhibit a highly variable and complex organization (Schostak *et al*., 2008). They encode reverse transcriptase and endonuclease from their single ORF that is different from the LTR and non-LTR retrotransposons (Evgen'ev and Arkhipova, 2005). Their endonuclease is similar to the GIY-YIG motif-containing group I introns, as well as bacterial UvrC DNA repair proteins, whereas their reverse transcriptase resembles the RT domain of telomerase.

1.12 DNA based transposons

DNA based transposons were first discovered as a spontaneous insertion in bacteria which prevents transcription-translation of the target gene. These elements are present in both prokaryotes and

eukaryotes. They always end in internal inverted terminal repeats and generate short repeats in the target site after insertion. They encode transposase which is responsible for the transposition of these elements. In the process of transposition to a new site, they make staggered breaks in the target DNA followed by joining the element to the protruding single-stranded ends and filling in the gaps. This is responsible for the formation of direct repeats of target DNA at the site of insertion. DNA based transposition takes place by both replicative and nonreplicative mechanisms. In replicative transposition, the element duplicates during the reaction; one copy inserts at the new location and the other remains at the original site. Replicative transposition occurs via two enzymatic activities: a transposase which works on the ends of the original transposon: and a resolvase which works on the duplicated copies (Craig, 2002). In non-replicative transposition, the element moves directly from the donor site to target site and is lost from the donor site. The donor site is then repaired by the host repair system.

1.13 Mechanism of non-LTR retrotransposition

Unlike the LTR retrotransposons or DNA transposons, non-LTR retrotransposons do not code for an integrase or transposase and use a different mechanism called target primed reverse transcription (TPRT) for their mobilization. The element encoded reverse transcriptase uses pre-existing breaks in the genome (or nicks introduced by the element encoded EN) as a primer to generate a cDNA copy of the RNA transcript directly onto the DNA break. The TPRT model is mainly based on studies with the R2 element of B. mori (Fig.7A and B), which inserts in 28S rRNA gene (Luan and Eickbush, 1995 and 1996; Luan *et al*., 1993), and binding properties of R2 protein to R2 encoded RNA (Christensen *et al*., 2006; Kurzynska-Kokorniak *et al*., 2007). The retrotransposition process starts with the transcription of R2 element to form mRNA which serves as both the template for translation of element-encoded protein and for retrotransposition. R2 RNA which contains a single ORF is translated into the R2 protein which contains N-terminal DNA binding domain, centrally localized RT domain and a C-terminal EN domain. R2 protein forms an RNP particle by binding near the 5'-and 3'-end of the R2 RNA, followed by its transportation to the nucleus where it binds to DNA target site. Two R2 protein subunits carry out the complete reaction of TPRT (Figure 7A). The R2 protein bound at the 3'-end of R2 transcript adopts a conformation that binds target DNA (28S gene) upstream of the insertion site. Whereas, R2 protein at 5'-end of the R2 transcript adopts a conformation that promotes binding to target DNA downstream of the insertion site. R2 subunit bound to the upstream site initiates retrotransposition reaction by nicking the bottom strand of target DNA and using the released 3'-OH to prime the reverse transcription reaction termed as TPRT. After completion of the first strand synthesis, the second R2 subunit that binds downstream to insertion site nicks the top DNA strand. The RT of the downstream subunit uses 3'-DNA end released by this cleavage to prime second strand DNA synthesis (Fig.7B). Once both DNA

strands are synthesized, the mRNA template must be removed. The RNase H domain which most likely removes the mRNA template has been acquired in some clades (Malik *et al*., 1999; Blesa *et al*., 1997). However, the R2 protein lacks an RNase H domain, thus it is not very clear how the mRNA template is removed. General features of TPRT model of non-LTR retrotransposition have also been supported by *in vivo* assays of L1 and I retrotransposition (Chaboissier *et al*., 2000; Cost *et al*., 2002).

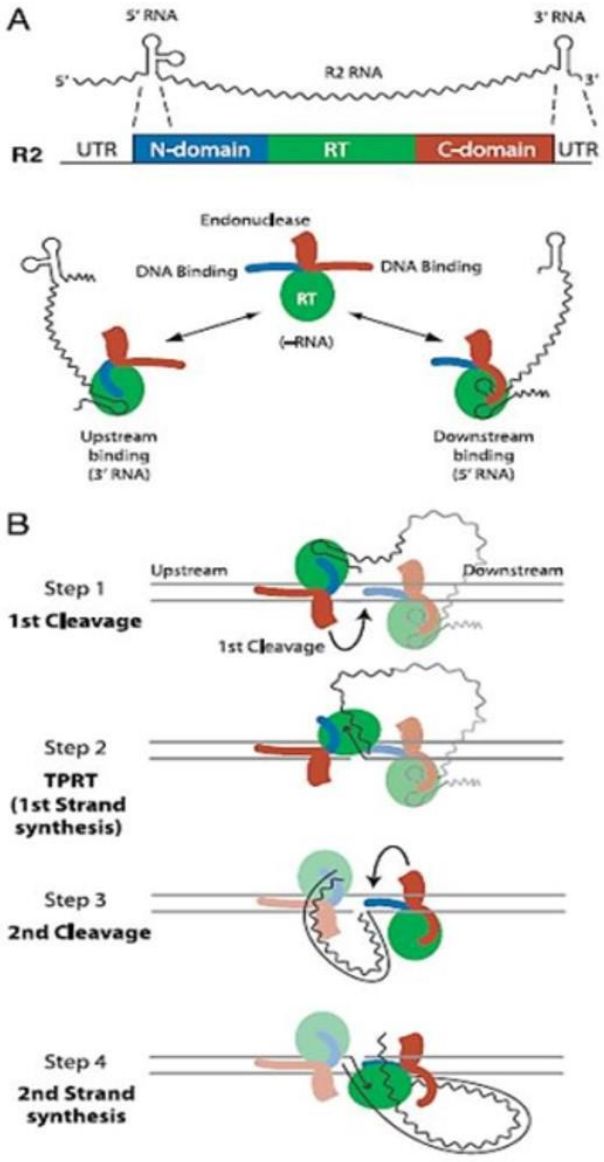


Figure 7. Target Primed Reverse Transcription (TPRT) Model for non-LTR retrotransposition (based on the R2Bm element). (A) R2 element encodes a single polypeptide containing three domains; a DNA-binding domain at N-terminal, a reverse transcriptase (RT) domain located centrally, and a DNA binding and endonuclease domain at C- terminal. R2 protein has the ability to identify the secondary structures at the 5'-and 3'-UTR of the R2 transcript. R2 protein association to the 3'-UTR of RNA sequesters the N-terminal DNA-binding domain (i.e., the upstream subunit conformation) and its association to the 5'-UTR of RNA sequesters the C-terminal DNA-binding domain (i.e. the downstream subunit conformation). (B) R2 retrotransposition initiates by two subunits

in four steps. In the first step, the upstream subunit derived endonuclease (EN) (red oval) cleaves the bottom strand of target DNA at the insertion site and leaves the 3'-OH free. The reverse transcriptase (RT) (green oval) from the upstream subunit catalyzes first-strand TPRT in the second step. In the third step, the EN derived from downstream subunit cleaves the upper DNA strand and polymerase performs the second-strand TPRT in the fourth step by displacing the RNA. The cleavage of upper DNA strand occurs only when the reverse transcription strips away the 5'-RNA region bound to the downstream subunit. (Adapted from Christensen *et al.*, 2006; Eickbush and Jamburuthugoda, 2008).

This model clearly explains two main features of non-LTR retrotransposon insertions. Frequently observed 5'-truncation of inserted element may arise due to RNA template degradation or incomplete reverse transcription. Also, this mechanism explains the generation of variable lengths of target site duplications (TSD) or in some cases target site deletion. Sequence complementarity of template RNA and target DNA is not required for priming in TPRT model. In case of deletions at the 3'-end of R2 mRNA sequence, R2 RT adds non-templated nucleotides (usually T nucleotides) to the target DNA (Luan and Eickbush, 1995). Thus, the addition of T nucleotides may explain the presence of poly (A) tails at the 3'-end of some non-LTR retrotransposons. However, it does not explain the presence of A-rich sequences or tandem copies of simple repeats (e.g., TAA) in some elements. So far, cleavage of the second DNA strand after reverse transcription has been shown only for the R2 element, but it can be inferred from the evidence of reduced cleavage or delayed kinetics of the second DNA strand by the endonucleases of other non-LTR retrotransposons (Anzai *et al.*, 2001; Christensen *et al.*, 2000; Feng *et al.*, 1996; Feng *et al.*, 1998). Second strand cleavage can occur upstream, downstream, or in line with the bottom strand nick based upon the type of non-LTR retrotransposon and generates target site deletions, target site duplications (TSDs), or blunt insertions respectively.

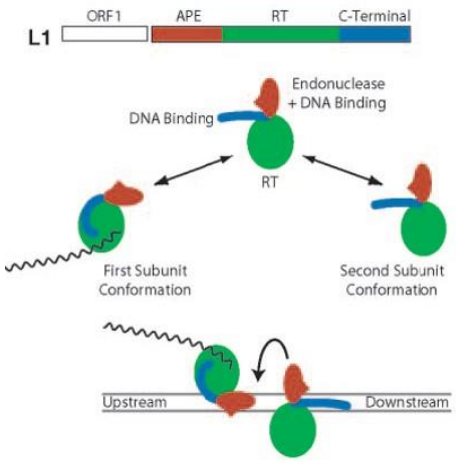


Figure 7C. Model for insertion of the human L1 element. L1 elements encode two ORFs, ORF1 and ORF2. The ORF2 contains an N-terminal AP endonuclease domain (red), a reverse transcriptase domain (green), and a C-

terminal domain (blue). As in the R2 model, the active complex is assumed to be a dimer, with each subunit in opposite orientation conducting one-half of the reaction. One subunit binds RNA, then binds the target DNA (upstream binding) by means of the APE domain, cleaves the first strand, and conducts TPRT. The second subunit binds to target DNA (downstream binding) by means of the C-terminal domain, cleaves the second strand, and conducts second-strand DNA synthesis. The role of ORF1 in this model could not be confirmed. There is no direct evidence of binding of the C-terminal domain of an APE-containing element to downstream of the insertion site. (Adapted from Christensen and Eickbush, 2005).

Studies with human L1 ORF2 also provide evidence for the TPRT reaction (Fig.7C). Unlike the R2 ORF, L1 ORF2 does not nick a rigidly defined target DNA. Ancient non-LTR clades, such as CRE, R2, and R4, possess REL ENDO type endonuclease that recognizes specific sequences. The REL ENDO type domain was substituted by an AP type endonuclease early in non-LTR evolution (Malik *et al*., 1999) and caused relaxed target site specificity. Though EN domain has an important role in TPRT reaction, EN-independent L1 retrotransposition has also been reported in the mammalian system (Morrish *et al*., 2007). The ORF2p also functions in the synthesis of second DNA strand primed from the DNA target. Besides retrotransposition of *cis* elements (L1 RNA) (Wei *et al*., *Introduction* 31 2001) L1 machinery also retrotransposes *trans* elements (SINEs and processed pseudogenes) (Dewannieux *et al*., 2003; Esnault *et al*., 2000).

1.14 Transposons in protozoan parasites

Transposable elements have been reported in many protozoan parasites including Trypanosomes, *Giardia*, *Crithidia*, *Leishmania*, and *Entamoeba*. Similar to eukaryotic genomes, these protozoan parasites have been colonized by diverse repetitive elements (Bhattacharya *et al*., 2002; Wickstead *et al*., 2003). Of all classes of transposable elements, non-LTR retrotransposons appear to be the most abundant in parasitic protozoa (Bhattacharya *et al*., 2002).

1.14.1 Non-LTR retrotransposons in Trypanosomes

Trypanosomatids include protozoan parasites which are of medical and veterinary significance. They lead to serious diseases in humans, such as sleeping sickness (*T. brucei*), Chagas disease (*T. cruzi*) and Leishmaniasis (*Leishmania* spp.). Retrotransposons in *T. brucei* and *T. cruzi* are of similar type and both species do not contain DNA transposons. All the trypanosomatid non-LTR retrotransposons belong to either Ingi or CRE clades (Eickbush and Malik, 2002) of which Ingi clade is the most abundant. On the basis of RT domain, Ingi clade has been divided into three subclades namely L1Tc (*T. cruzi*), LmDIRE (*L. major*), and Ingi (*T. brucei and T. cruzi*) (Bringaud *et al*., 2009). The ingi subclade is further divided into three groups; TbDIREs (*T. brucei*), TcDIREs

(*T. cruzi*) and Tbingi. The DIREs ("degenerate *ingi*/L1Tc-related elements") are highly degenerate elements and contain various frame-shifts and stop codons (Ghedin *et al*., 2004). The Tbingi (5.25kb) and L1Tc (4.74kb) are potentially functional and autonomous retrotransposons. L1Tc is a well-characterized element and encodes three ORFs (Martin *et al*., 1995); ORF1 polypeptide shows similarity with AP endonuclease and encodes a protein with AP endonuclease activity *(Olivares et al*., 1997), ORF2 shows homology with RT domain of non-LTR elements and ORF3 shows the presence of two motifs ($CX_2CX_{12}HX_{3\text{-}5}H$) similar to the CCHH class of zinc finger. Same motifs are present in all the trypanosomatid non-LTR retrotransposons and in the insect R2Bm element. The trypanosome genome also shows the presence of small nonautonomous retroposons, namely, NARTc (0.26kb) and 0.5kb long TbRIME which are related to the autonomous L1Tc and Tbingi respectively (Bringaud *et al*., 2002; Hasan *et al*., 1984). Tbingi/TbRIME, L1Tc/NARTc, and DIREs share the first 79 residues at their 5'-ends, that constitutes the hallmark of trypanosomatid retroposons ("79-bp signature"). In addition to above-mentioned species of trypanosomatids, the DIREs have also been characterized in *T. congolense*, *T. vivax*, and *L. braziliensis*. Small degenerate retroposons (0.55kb) containing the "79-bp signature," named LmSIDERs (for *s*hort *i*nterspersed *de*generate *r*etroposons), have also been identified in the genomes of *L. major* (Bringaud *et al*., 2007), *Leishmania infantum*, and *L. braziliensis* (Smith *et al*., 2009). SIDER elements are quite uniformly dispersed throughout all these three genomes. There is evidence for the species-specific enrichment of SIDERs and for their preferential association, especially for SIDER2s, with different metabolic functions. Evolutionary relationship of SIDERs to other trypanosomatid retroposons reveals that SIDER1 is a truncated version of extinct autonomous *ingi* like retroposons (DIREs), which were functional in the ancestral *Leishmania* genome (Smith *et al*., 2009). The initial characterization of Tbingi/TbRIME and L1Tc/NARTc retroelements suggested their random distribution in *T. brucei* and *T. cruzi* genomes. But genome sequence has revealed that these retroelements are not randomly distributed. A large multigene family called RHS (retrotransposon hot spot) which contains a hot spot for insertion has been characterized in *T. brucei*. Analysis of these hot spots revealed the presence of a conserved sequence (-34 Axxxxxxx**<u>TtgxTGxGGxT</u>**xxx **tTxTxT** -6) upstream of the Tbingi/TbRIME retroelements [where x denotes any nucleotide], with an 11bp core consensus sequence (underlined residues) located 4-14bp upstream of the first single strand cleavage (Bringaud *et al*., 2004). Similarly, a well-conserved motif (**GA**xx**A**x**Ga**xxxxxtx**TATG↑A**xxxxxxxxxxx; the arrow indicates the first-strand cleavage site) precedes L1Tc/NARTc retroelements in *T. cruzi* (Bringaud *et al*., 2006). Another unique feature of these elements is the presence of 12 bp TSD flanking majority of these elements. Interestingly exactly 12 bp TSD has been found in Tbingi/TbRIME (Bringaud *et al*., 2004),

L1Tc/NARTc (Bringaud, 2005) and Tcoingi (*T. congolense*) and Tcingi (*T. vivax*) elements (Bringaud *et al*., 2009).

The CRE, second clade of trypanosomatid retrotransposons is composed of the SLACS (*T. brucei*), CZAR (*T. cruzi*) and CRE1/CRE2 (*Crithidia fasciculata*) elements *(Aksoy et al., 1990; Gabriel et al*., 1990; Villanueva *et al*., 1991). These are site-specific retroelements always inserted at the same relative position in the spliced leader (*SL*) *RNA* genes. SLACS element (6678bp) has two ORFs. ORF1 encodes a gag like polypeptide of retroviruses. ORF2 encodes a protein with RT and EN domains (Aksoy *et al*., 1990). A unique feature of SLACS is that all copies are conserved in sequence. There are no truncated copies and all have the same 49bp TSD. CZAR is a site specific non-LTR retrotransposon in *T. cruzi* genome. Like SLACS it inserts specifically into SL-RNA genes (Villanueva *et al*., 1991). Full-length CZAR elements are 7kb in length and contain two ORFs. Amino acid sequence comparisons indicate that, like SLACS, the CZAR ORF1 has a CCHH motif with nucleic acid binding properties. The CZAR ORF2-encoded protein has conserved reverse transcriptase and endonuclease domains. CRE1/CRE2 elements are described later.

1.14.2. Non-LTR retrotransposons in *Giardia lamblia*

It infects the small intestine of human, and variety of other mammalian hosts. Three families of non-LTR retrotransposons have been reported in *G. lamblia* (Arkhipova and Morrison, 2001). Two of these, GilM and GilT, are potentially active elements while the third, GilD, is probably composed entirely of inactive copies. Both GilM and GilT are confined to immediate subtelomeric regions in tandem head to tail orientation. Members of such a tandem array are separated from each other by the $(A)_n$ stretch (n = 10-16). The coding region of GilT and GilM consists of a long ORF, about 1000 amino acids in length, which is preceded by a short 5'-UTR (55bp). The ORF consists of a central RT domain, N-terminal zinc finger motif of the CCHH type and C-terminal REL-endonuclease domain.

1.14.3 Non-LTR retrotransposons in *Crithidia fasciculata*

As mentioned above, two retroelements (CRE1 and CRE2) have been reported in *C. fasciculata* genome. Both elements are sequence specific, inserting specifically in the mini exon gene locus (SL). CRE1 has a single ORF of 3420 nucleotides in length. It ends with a poly(A) stretch which varies from 16-57 nucleotides (Gabriel *et al*., 1990). Conceptual translation of CRE1 ORF reveals that it can encode a protein containing central reverse transcriptase domain, an N-terminal nucleic acid binding domain, and C-terminal REL endonuclease domain. CRE2 is 9.6kb long, has a single ORF, predicted to encode a protein containing reverse transcriptase and endonuclease domain like

CRE1 (Teng *et al*., 1995). CRE1 and CRE2 have approximately 30% identity over a 1000 amino acid region towards the C-terminus of the ORF. Beyond this, the two elements are structurally distinct. Whereas CRE2 has an 844bp 5'-UTR, CRE1 has no apparent 5'-UTR. Therefore two evolutionarily diverged retrotransposons share the same insertion site within the same genome (Bhattacharya *et al*., 2002).

1.14.4 Non-LTR retrotransposons in *E. histolytica*

E. histolytica contains three families of non-LTR retrotransposons and very few DNA transposons. Conversely, four DNA transposon superfamilies; *hAT*, *Mutator*, *piggyBac*, and Tc1/*mariner* have been identified in *E. invadens* and *E. moshkovskii* which have very few retrotransposons. Only the mutator family of DNA transposons is found in *E. histolytica* with very few copies (Pritham *et al*., 2005). Historically, the presence of non-LTR retrotransposons in *E. histolytica* was shown first from our laboratory by BLASTX analysis of a multicopy DNA sequence, called HMc. It showed significant similarity with the RT sequence of non-LTR retrotransposons (Sharma *et al*., 2001; Mittal *et al*., 1994). Further analysis showed that HMc was part of a large sequence with similarity to known non-LTR retrotransposons and was termed EhRLE (*E. histolytica* retrotransposon like element). Subsequently, completion of *E. histolytica* genome project (Loftus *et al*., 2005) revealed sequences of all copies and the EhRLE was later termed EhLINE1. Genome sequence analysis of *E. histolytica* showed the presence of multiple autonomous and nonautonomous non-LTR retrotransposon families now designated EhLINEs and EhSINEs respectively (Bakre *et al*., 2005; Van Dellen *et al*., 2002). With the help of *E. histolytica* genome reassembly and reannotation along with complete genome sequencing of *E. invadens* and *E. dispar*, all the families of TEs have been well identified computationally in this protozoan parasite. It has been estimated that in *E. histolytica*, repetitive sequences represented about 19.7% whereas non-LTR retrotransposons (LINEs and SINEs) account for 11.2% of the genome. Novel families of TEs known as EREs (*Entamoeba* Repetitive Elements) have been identified in addition to EhLINEs and EhSINEs (Lorenzi *et al*., 2008). ERE1 was found to be present in all three *Entamoeba* species while ERE2 was detected only in *E. histolytica* (table 2) (Lorenzi *et al*., 2008). These elements shared some common features like AT richness, presence of single ORF, two TIRs (terminal inverted repeats) and their integration into the intergenic regions. It has been proposed from the EST database that there is evidence of transcription of ERE1 and ERE2 in this parasite.

LINEs and SINEs are classified into three families (EhLINE1, EhLINE2, EhLINE3 and their non-autonomous partner EhSINE1, EhSINE2 and EhSINE3). EhLINE1 (4.8kb) is the most abundant

family of TEs in *E. histolytica* with 88 complete copies and a total of 742 elements (Lorenzi *et al*., 2008) (Table 2). *Entamoeba dispar* is a sibling species of *E. histolytica*, is found in the human gut, but is nonpathogenic. It also contains three families of LINEs and SINEs. EdLINE1 (4.8kb) represents the most abundant family of LINEs with a total of 573 copies with 63 complete copies. There are 442 copies (73 complete) of EhLINE2 as compared to 449 (28 complete) copies of EdLINE2 whereas 87 copies (10 complete) of EhLINE3 and 42 (2 complete) copies of EdLINE3. In *E. invadens* only one LINE family (EiLINE1) is found. EiLINE1 has 67 copies with only 2 complete copies. SINEs, the non-autonomous partners of LINEs, are also present in three families in both *E. histolytica* and *E. dispar*. EhSINE1 (550bp) is the most abundant family having 445 copies (264 complete) followed by EhSINE2 and EhSINE3 having 256 (94 complete) and 49 (9 complete) copies respectively in *E. histolytica*. EdSINE1 is the most abundant with 425 (282 complete) copies followed by EdSINE2 and EdSINE3 with 189 (53 complete) and 18 (2 complete) copies respectively in *E. dispar* (Table 2). On the basis of sequence analysis of SINEs in *Entamoeba,* it indicates that EhSINE3/EdSINE1 existed as a chimeric element in common ancestor of *E. histolytica* and *E. dispar*. The 5'-end of this chimeric element was derived from precursor sequence of EhSINE2/EdSINE2 and the 3'-end was derived from ancestral EhSINE1-like element (Lorenzi *et al*., 2008).

name	complete[a]	incomplete[a]	coverage (bp)	coverage (%)[b]
Eh_LINE1	88	654	1079630	5.2%
Eh_LINE2	73	442	759971	3.7%
Eh_LINE3	10	87	160940	0.8%
Eh_SINE1	264	181	187972	0.9%
Eh_SINE2	94	162	108195	0.5%
Eh_SINE3	9	40	18851	0.1%
Eh_ERE1	0	777	1014754	4.9%
Eh_ERE2	71	728	733987	3.5%
Eh_MuDR	0	4	2851	< 0.1%
Eh_mariner	0	1	1008	< 0.1%
Eh_TOTAL	**609**	**3047**	**4068159**	**19.7%**
Ed_LINE1	63	510	839108	3.7%
Ed_LINE2	28	449	506244	2.2%
Ed_LINE3	2	42	46734	0.2%
Ed_SINE1	282	143	208941	0.9%
Ed_SINE2	53	136	73091	0.3%
Ed_SINE3	2	16	2497	< 0.1%
Ed_ERE1	51	536	526451	2.3%
Ed_MuDR	0	4	2075	< 0.1%
Ed_mariner	0	1	1011	< 0.1%
Ed_TOTAL	**481**	**1837**	**2206152**	**9.7%**
Ei_LINE	2	67	59,308	0.1%
Ei_ERE1	30	227	170,510	0.4%
Ei_DDE	328	2607	1,678,976	4.1%
Ei_mariner	390	1400	822,878	2.0%
Ei_hAT	35	755	464,161	1.1%
Ei_MuDR	49	831	522,116	1.3%
Ei_Polinton	5	126	336,005	0.8%
Ei_piggyBac	14	32	27,894	0.1%
Ei_TOTAL	**677**	**6082**	**4033163**	**9.9%**

[a] Repeats were considered complete when their length was at least 90% of the consensus sequence.
[b] Expressed as percentage of the corresponding genome length.

Table 2: Number and coverage of transposable elements in *E. histolytica, E. dispar and E. invadens* (Adopted from Lorenzi et al., 2008).

EhLINEs showed similarity with R4 clade elements e.g. Rex6 (from *Oryzias*), Dong and R4 (from *Bombyx mori*) and have similar domain structure with these elements. All these elements contain a putative nucleic acid binding motif (CCHC) and REL- ENDO type endonuclease located at the downstream of RT domain. The C_2H_2 zinc finger motifs which are found N-terminal to the RT domain in NeSL-1, R2, SLACS, and GilM, are lacking in these elements (Malik and Eickbush, 2000; Arkhipova and Morrison, 2001). Phylogenetic analysis on the basis of manually reconstructed consensus sequence of reverse transcriptase showed that LINEs found in *Entamoeba* species were derived from a single ancestral sequence that was present before they diverged from each other (Lorenzi et al., 2008).

Most copies of EhLINEs are truncated at their 5'-end or 3'-end or at both the ends in *E. histolytica*. Of the full-length copies, none was found to contain complete ORF (Bakre *et al*., 2005). However, some LINE1 and LINE2 elements in *E. histolytica* and *E. dispar* were identified that encoded either a complete ORF1 or a complete ORF2. Though most of the copies have multiple mutations, entries in the *E. histolytica* GSS (genome sequence survey) data base (which were used for *E. histolytica* genome sequencing) showed complete ORFs for the length of the GSS clone which was 600-800bp per read. With the help of GSS clones containing continuous ORFs, a copy containing consensus sequence of EhLINE1 was constructed (Bakre *et al*., 2005) and used for functional analysis. Southern hybridization of the EhLINE1 probe with PFGE separated chromosomes of *E. histolytica* showed that the element is present in all chromosomes of *E. histolytica*, and is not telomerically located (Sharma *et al*., 2001). Analysis of sequences flanking the insertion site of EhLINE1 and EhSINE1 resulted in short target site duplications (TSD) upon insertion. The sequences upstream and downstream of the EhLINE1 and EhSINE1 showed that these elements do not insert in a strictly site-specific manner however, a short T-rich stretch was found to be present upstream of the insertion site of both the elements (Mandal *et al*., 2004).

Similar to EhLINE1, EhSINE1 is also dispersed throughout the genome (Sharma *et al*., 2001). EhSINE1 shares 3′ ~74bp sequence with EhLINE1, and could thus be a partner SINE of EhLINE1 (Fig.8). EhSINE1 is not a tRNA-derived SINE as its sequence does not show any match with *E. histolytica* tRNA genes. The origin of EhSINEs is not clear as they do not show homology with any known sequences.

1.14.4.1 Proteins encoded by EhLINE1

Consensus sequence analysis showed that EhLINE1 is 4804bp in length with short 5'-and 3'-UTRs and consists of two non-overlapping ORFs, ORF1 and ORF2 (Fig.8). The two ORFs have different reading frame and are separated by a 443bp non-coding spacer (Bakre *et al*., 2005). ORF1 encodes a polypeptide of 498 amino acids (nucleotide position 14 to 1510). The orf1 polypeptide of most non-LTR retrotransposons has cys-his motif which is present in the proteins that bind to nucleic acids. EhLINE1 does not contain cys-his motif but it has homology with nucleic-acid-binding proteins and those mediating protein-protein interactions. Although EhLINE1-encoded ORF1 lacks easily recognizable functional domains, the possible functional domains in EhLINE1-ORF1p have recently been predicted in our lab bioinformatically. It showed the presence of a long RNA-binding stretch at the N-terminal and some small stretches throughout the sequence. In addition, a C-terminal coiled coil domain which participates in protein-protein interaction was also predicted

along with two nuclear localization signals (NLS) in EhLINE1-ORF1p (Gaurav *et al*., 2017). Biochemical analysis showed that EhLINE1-ORF1p lacked sequence-specificity of RNA-binding, and could bind to EhLINE1-RNA fragment and ribosomal RNA with comparable affinities.

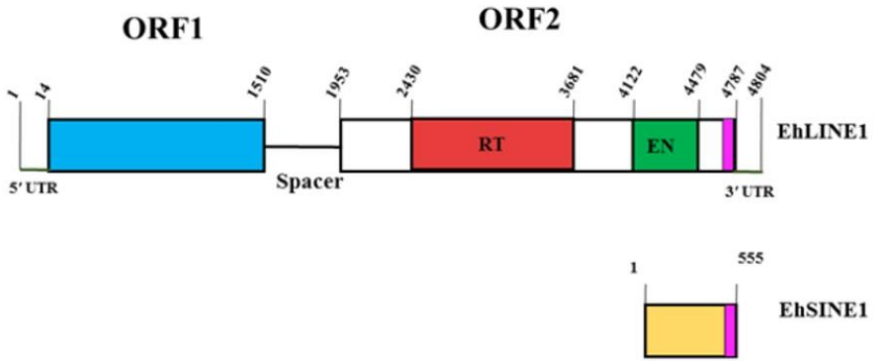


Figure 8: Schematic representation of EhLINE1 and EhSINE1. EhLINE1 is 4804bp long with short 5'-UTR (14 bp) and 3'-UTR (17bp). It contains two non-overlapping ORFs, ORF1 and ORF2, separated by a 443bp spacer region. ORF1 has a nucleic acid binding activity and ORF2 contains a central RT and C-terminal EN domain. EhSINE1 is 555bp long and shares a 74bp sequence similarity with EhLINE1 at the 3'-end (shown in the pink bar). (Adopted from Bakre *et al*., 2005)

ORF2p contains a central reverse transcriptase domain (RT) (nucleotide position 2430 to 3681) and a C-terminal REL-ENDO domain (EN) (nucleotide position 4122 to 4479) (Fig.8). The RT domain has the closest match with the RTs encoded by R4 clade of non-LTR retrotransposons, most notably the R4 element of *Ascaris lumbricoides* and the Dong element of *Bombyx mori*. The RT domain has a highly conserved motif, YXDD which is conserved through retroviruses, LTR, and non-LTR retrotransposons. Lentiviral RTs contain methionine residue at the place of "X" in this conserved motif which is thought to be responsible for the low fidelity of these RTs (Kaushik *et al*., 2000; Poch *et al*., 1989). The EhLINE1 ORF2 also has methionine at the same location in RT. The R2 group of non-LTR retrotransposons has highly conserved CCHC, $PDX_{12\text{-}14}D$, RHD and KXXXY motifs in EN domain which is also found in EhLINE1 (Yang *et al*., 1999). EhLINE2 and EhLINE3 show the sequence organization similar to EhLINE1 and have well conserved RT and EN domains. The enzymatic properties of the recombinant EN domain have been well characterized in our lab (Mandal *et al*., 2004; Yadav *et al*., 2009). The EN activity has been studied with a 176bp DNA substrate derived from flanking sequences of a genomic insertion site of EhSINE1. This sequence was selected because *E. histolytica* genome contains both an "empty" site and "occupied" site of the same sequence. The ENp nicked the bottom strand of 176bp DNA substrate at three hot spots; one

of these hot spots was the exact genomic insertion site for EhSINE1 (Mandal *et al*., 2004). Despite being REL-ENDO type endonuclease, the EhLINE1 encoded EN lacked strict sequence specificity in target-site recognition. It could possibly recognize structural features of the DNA as well. With a variety of substrates derived from the 176bp DNA, the sequence specificity of EN was checked and a consensus sequence (5′-GCA↑T↑T-3′, arrows denote the nicking sites) where the enzyme nicks most frequently, was assigned (Mandal *et al*., 2006). *In vivo* retrotransposition was studied in ORF2-expressing *E. histolytica* cell lines (since ORF1p is constitutively expressed) (Yadav *et al*., 2012). To study i*n vivo* retrotransposition, a marked copy of EhSINE1 (with a GC-rich tag) was expressed constitutively. The same vector also contained 176 bp target DNA, along with ORF2 cloned in a tetracycline inducible vector. Upon induction of ORF2 expression by adding tetracycline, retrotransposition of marked-EhSINE1 copy was monitored by scoring its insertion into the 176bp target DNA. Sequence comparison of the newly retrotransposed EhSINE copies showed the appearance of chimeric EhSINEs at high frequency. These chimeric EhSINEs contained sequences from both genomic SINEs and the marked-SINE. These observations were hypothesized to be the product of template switching property of RT of ORF2p during retrotransposition.

1.15 Characteristics of non-LTR retrotransposons

1.15.1 LINE1 transcript studies

Out of the 500,000 copies of L1 elements, only a few thousand are full length and contain the internal 5'-UTR promoter which can generate the RNA species capable of retrotransposition. These retrotranspositionally competent as well as defective, full-length and truncated L1s are spread throughout the genome. Many copies are found to be present in introns and 3'-noncoding regions of cellular genes. Hence, it may result in the partial incorporation of L1 elements into the transcripts of other genes. LINE repeats accumulated in primate and rodent genomes are mostly retrotranspositionally inactive but some of these can be expressed. Young L1 subfamilies are capable of expression and retrotransposition (Lander *et al*., 2001; Waterston *et al*., 2002). A complex variety of RNAs are produced by primate L1 loci (Belancio *et al*., 2006; Perepelitsa-Belancio *et al*., 2003; Nigumann *et al*., 2002; Speek, 2001), some of which have capability to retrotranspose and others represent “deadend” transcripts with a yet unknown function in L1 amplification or cellular fate. The complexity of primate L1 transcripts is partly due to the presence of multiple transcription start sites. In human L1 5'-UTR at least two regions of sense start sites (Alexandrova *et al*., 2012; Swergold, 1990; Athanikar *et al*., 2004) and an antisense start site (Speek, 2001) have been identified. The full-length L1 mRNA transcripts are generated from the

entire length of the L1 locus by the sense L1 promoter which is present within the first 100bp of the L1 5'-UTR (Swergold, 1990). Some human L1 transcripts show 5'-truncation and absence of much of the 5'-UTR due to transcription initiation near the downstream end of the L1 5'-UTR (Alexandrova *et al*., 2012). It may be possible that both the full-length and 5'-truncated transcripts are generated from the same promoter, but in one case the polymerase reaches upstream and in other downstream of the promoter. Truncated copies have been shown to affect the process of retrotransposition. Over a thousand human L1 loci, containing stop codon in their ORF1 sequence, have been revealed bioinformatically. These loci showed expression as confirmed by RNA Seq analysis and resulted in the generation of truncated ORF1 proteins. These truncated ORF1 proteins suppress human L1 retrotransposition in trans provided that coiled-coil (CC) domain remained intact, as a mutation within CC domain abolished the suppressive effect of truncated proteins on L1 retrotransposition (Sokolowski *et al*., 2017). 5'-truncated L1 mRNAs can be expected to be able to mobilize provided that they encode both the L1 proteins (Belancio *et al*., 2006; Moran *et al*., 1996). Mouse L1 subfamilies also generate a complex set of RNAs as suggested by available evidence (Belancio *et al*., 2006; Perepelitsa-Belancio *et al*., 2003).

RNA processing has been demonstrated as the major regulator of L1 activity. Processing of L1 produces translatable spliced transcript (SpORF2) in many tissues which support expression of the functional ORF2 protein and induces DNA damage in normal cells (Belancio *et al*., 2010). L1 has also been shown to contain multiple functional canonical and noncanonical polyadenylation signals which result in the generation of 3'-truncated L1 transcripts due to premature polyadenylation (Perepelitsa-Belancio *et al*., 2003). Chimeric transcripts are known to be generated from L1 loci which may contain full-length or partial L1 sequences joined with the adjacent or distant genomic sequences (Belancio *et al*., 2006; Nigumann *et al*., 2002; Speek, 2001; Matlik *et al*., 2006) (Fig.9). In addition to chimeric transcripts, L1 sequences can also join with distant genomic sequences by utilizing the splice donor and acceptor sites that are distributed throughout the L1 sequence (Belancio *et al*., 2006; Nigumann *et al*., 2002).

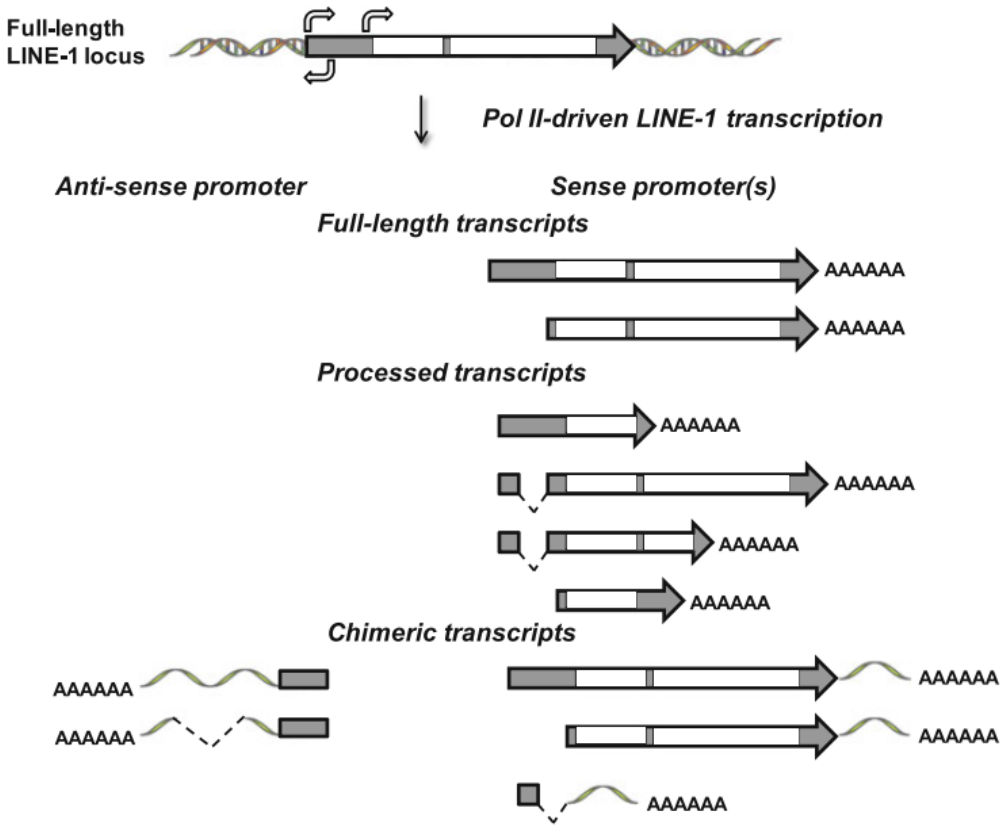


Figure 9: A schematic picture of different L1 transcripts generated from a full-length transcriptionally active L1 locus. Horizontal arrows above and below the L1 represent relative positions of sense and antisense transcription start sites, respectively. Dashed lines show spliced mRNAs while wavy lines represent genomic sequences adjacent or distant to L1 locus (Adopted from Deininger *et al.*, 2016)

1.15.2 Bidirectional transcription in LINEs and its implication on gene regulation

Most mammalian retrotransposons remain transpositionally incompetent due to several mutations and truncations. However, this does not exclude the ability of transcription initiation from promoters present within the elements (Conley *et al.*, 2008). Human and mouse L1 elements have been shown to contain a sense and antisense promoter (ASP) in the 5'-UTR, resulting in bidirectional transcription. There is also a promoter in the 3'-UTR which is in the sense orientation and can read outward into the flanking genomic sequence (Faulkner *et al.*, 2009). Promoter bidirectionality can generate a large number of antisense transcripts including chimeric transcripts, non-coding RNA, antisense mRNA or double stranded RNA (dsRNA), which affects the gene expression (Nigumann *et al.*, 2002) (Fig.10). Antisense transcription, which was earlier considered as transcriptional noise is now known as an important gene regulator. It can work as a regulatory switch by rewiring the regulatory network. The arrangement of antisense RNAs against sense genes

in the form of double stranded RNA (dsRNA) allows genes to regulate their own expression through self-regulatory circuits.

The ASPs of transpositionally incompetent L1s have been shown to function as alternative promoters for various human protein-coding genes (Speek, 2001; Matlik *et al*., 2006; Nigumann *et al*., 2002). cDNA mapping and RNase protection assay demonstrated that ASP of human L1 is directly involved in the transcription of adjacent cellular genes, producing chimeric transcripts which contained L1 5'-UTR along with cellular mRNA (speek, 2001). In addition, EST analysis for the mRNAs initiating in the L1 5'-UTR showed that L1 ASP affects the cellular transcriptome by generating putative L1 antisense chimeric transcripts that appeared to be alternative genic transcripts (Criscione *et al*., 2016).

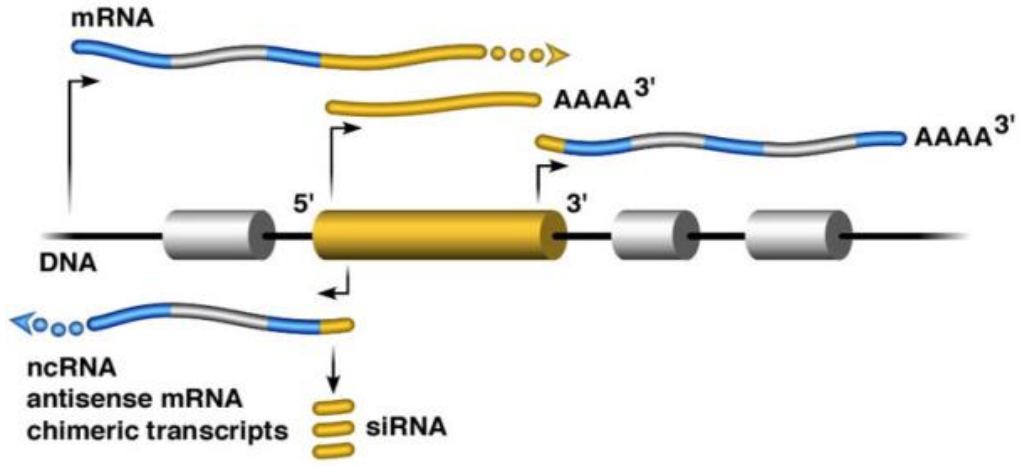


Figure 10: Sense and antisense promoter effect: Human L1 elements shows the presence of sense and antisense promoter within their 5'-UTR and another promoter in the 3'-UTR (bent arrows). The promoter in their 5' and 3'-UTRs are able to initiate ectopic transcription of L1 flanking sequences. ASP activity may generate ncRNA, antisense mRNA or chimeric transcripts. Further, ASP activity can reduce mRNA levels through the formation of double-stranded RNA (dsRNA) which lead to the production of siRNA that induces RISC-dependent silencing (Yang *et al*., 2006) (Adopted from Singer *et al*., 2010).

Antisense transcription is a widespread (Katayama *et al*., 2005) and well-characterized mechanism with the help of which ncRNAs are known to modulate the epigenome. Bidirectional retrotransposon transcription can also characterize chromatin boundaries such as SINE B2 which acts as a boundary element and maintains euchromatin around the mouse growth hormone locus (Lunyak *et al*., 2007).

1.15.3 Polyadenylation of LINE transcripts

Non-LTR retrotransposons are also called poly(A) retrotransposons due to the presence of a stretch of "A" residues at their 3'-end. The majority of the L1s cannot mobilize as they are 5'-truncated or have accumulated mutations (Grimaldi *et al*., 1984). The elements which have lost the function due to truncation may still have polyadenylation sites which can regulate the transcription of a genomic region (Mourie *et al*., 2008; Wheelan *et al*., 2005). Generally, poly(A) site is composed of the three main cis-acting elements: a conserved hexamer (AATAAA), cleavage site, and a GU-rich downstream region. An L1 copy which is inserted to the new genomic location contains the conserved AATAAA hexamer at its 3'-end. The GU-rich downstream region has to be provided by the neighboring genomic sequences (Belancio *et al*., 2007). This downstream region can dictate the strength of the poly(A) site depending on its affinity to cellular factors (Legendre and Gautheret, 2003). L1 encoded proteins (ORF1p and ORF2p) that preferentially mobilize the L1 RNA in cis can also mobilize Alu RNA and, more rarely, cellular mRNAs in trans (fig.11) (Doucet *et al*., 2015). Although these RNA differ in their sequences, they contain a common 3'-poly(A) tail which is crucial for the process of retrotransposition (Boeke, 1997; Richardson *et al*., 2015).

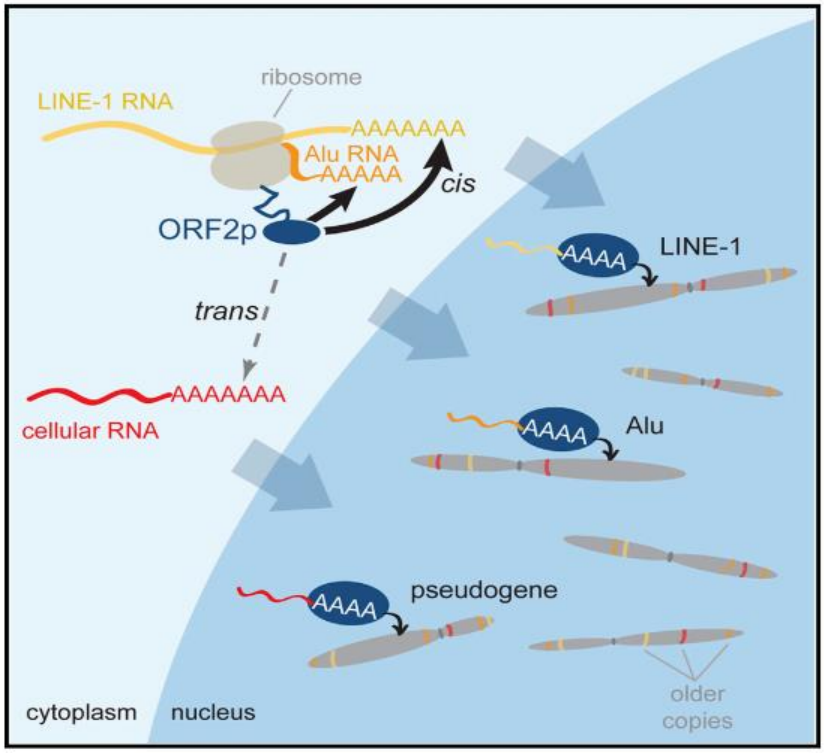


Figure 11: A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. L1 retrotransposons encoded ORF2p preferentially associate with their encoded RNA and Alu RNA in cis and cellular mRNAs in trans. Such association leads to the assembly of a ribonucleoprotein particle (RNP) and gets transported to the nucleus for the retrotransposition (Adopted from Doucet *et al*., 2015).

In Alu RNA a 50-base poly(A) tract at the 3'-end is required for efficient Alu retrotransposition (Dewannieux *et al*., 2003; Dewannieux and Heidmann, 2005). There are several sites in LINE1 which are predicted to be stronger than the poly(A) site at the 3'-end. In L1, polyadenylation signal has often been described as weak (Moran *et al*., 1999; Pickeral *et al*., 2000), suggesting that the 3'-flanking sequences may have the ability to alter the size and quantity of the L1 transcripts produced from different loci. The presence of internal poly(A) signals suggests that they have a conserved function either in limiting or regulating the LINE-1 retrotransposition. Northern blot analysis detected abundant truncated bands that corresponded in size to the positions of the putative poly(A) signals which show their functional relevance in mammalian LINE-1 elements (Perepelitsa-Belancio *et al*., 2003).

1.16 DNA methylation and transcriptional repression in Retrotransposons

DNA methylation is an important epigenetic regulation and is correlated with transcriptional silencing, genomic imprinting and cellular differentiation (Bird, 2002; Reik, 2007). For the viability of any species, genome should be kept stable and should be transmitted from one generation to the next through germ cells. In addition to function at gene promoters, DNA methylation also plays role in silencing of TEs and thereby reducing their threat to genomic integrity (Reik, 2007). The majority of TEs are inactive due to gene degeneration and some active TEs are silenced by DNA methylation (Bourc'his and Bestor, 2004; Kaneda *et al*., 2004). The mutational activity of transposable elements, particularly retrotransposons, interfere with the genome stability. Transcriptional silencing mechanism in retrotransposons, mediated by methylation, have been studied in germ cells where retrotransposons are activated and can integrate into new genomic locations and influence the nearby gene expression by recombination, rearrangement and mutation in the host genome (Romanish *et al*., 2010). However, germ cells have evolved defense mechanisms which suppress retrotransposon function and maintain genome stability (Öllinger *et al*., 2010; Zamudio *et al*., 2010; Rowe *et al*., 2011). Mammalian cells use transcriptional silencing mechanism against the retrotransposons which typically involves addition and removal of covalent modifications to DNA and histones (Cedar *et al*., 2009) and alter the chromatin structure.

DNA methylation is the modification which can develop heritable changes in gene activity without changing the DNA sequences. In mammals, DNA methylation mostly occurs at the cytosine residue in the context of CpG dinucleotides and takes place by the methyltransferases DNMT3A and DNMT3B and is maintained by the DNMT1 (Reddington *et al*., 2013; Li *et al*., 1992; Okano *et al*., 1999). Repetitive sequences are also methylated like most of the mammalian genome (Lister *et al*.,

2009; Meissner *et al*., 2008). Almost half of the CpGs in human genome reside in repeat sequences with 25% of them in Alus and 12% in LINEs (Rollins *et al*., 2006; Xie *et al*., 2009). In mammals, DNA methylation is strongly enriched at retrotransposons. Hence, the primary role of methylation has been proposed to be transcriptional repression of these elements (Yoder *et al*., 1997) and could exert its repressive effect by interfering with transcription factor binding sites in regulatory regions (Wiench *et al*., 2011). In general, a gene shows expression when most of the CpG islands in gene promoters are undermethylated. In somatic cells, L1 5'-UTR is heavily methylated and results in suppression of L1 expression (Crowther *et al*., 1991; Thayer *et al*., 1993; Hata *et al*., 1997; Woodcock *et al*., 1997). A number of methyl-CpG binding proteins have been identified in mammals such as MeCP2 which is thought to have a role in transcriptional repression through its interaction with SINE3A co-repressor complex (Nan *et al*., 1998). Although it is not confirmed if the methyl-CpG binding proteins are necessary for global silencing of target genes or retrotransposons in mammals (Caballero *et al*., 2009). However, it has been shown that experimental removal of DNA methylation induces IAP (LTR retrotransposon) expression in mouse embryonic fibroblast (Davis *et al*., 1989; Hackett *et al*., 2012; Jackson-Grusby *et al*., 2001). MMVL30 and MuLv, an LTR retrotransposon, are also an example which is transcriptionally upregulated in response to hypomethylation in mouse embryonic fibroblast (Brunmeir *et al*., 2010). Like the LTR retrotransposons, non-LTR retrotransposons LINEs and SINEs also show methylation in somatic tissues (Popp *et al*., 2010). However, transcriptional silencing of these elements does not depend strongly on fibroblast methylation (Brunmeir *et al*., 2010). DNA methylation and MeCP2 have been shown to be involved in human and mouse LINE-1 retrotransposons in neuronal cells (Muotri *et al*., 2010). Bisulfite sequencing at genome level showed that the non-LTR (LINEs and SINEs) and LTR (eRv1, eRvK, eRvL and MaLR) retrotransposons showed only 20% of their CpG methylation in germ cells (in which they are active) as compared to somatic tissue (80%) (Popp *et al*., 2010) whereas IAP elements showed ~60% of DNA methylation in germ cells (Seisenberger *et al*., 2012; Guibert *et al*., 2012; Lane *et al*., 2003).

Although promoter DNA methylation has been linked to transcriptional repression, in many cases it is not clear whether methylated promoters represent a cause or is a consequence of gene silencing (Walsh and Bestor, 1999; Bird, 2002). The ‘genome defense’ hypothesis suggests that the primary function of DNA methylation is silencing of TEs rather than regulation of developmental gene expression, and DNA methylation of CpGs evolved as a host defense mechanism against transposable elements (Yoder *et al*., 1997).

1.17 Aims and Objectives

Transcription of LINE element is the first step in the retrotransposition process. Only transcriptionally active copies are retrotranspositionally competent. Similar to other organisms, several mutations and truncations have been reported in EhLINE1 copies, with only a small fraction being full-length. The transcription status of the EhLINE1 family remains unanswered in *E. histolytica*. To answer this question, we undertook RNA Seq analysis to categories the EhLINE1/EhSINE1 copies on the basis of their expression level. Since methylation plays an important role in gene expression, determination of the extent of cytosine DNA methylation in EhLINE1 copies would help to understand the regulation of EhLINE1 expression. In addition, EhLINE1 ORF2p expression and enzymatic activity of encoded protein would help in the investigation of its involvement in RNP formation and retrotransposition as ORF2 plays an important function in retrotransposition via target primed reverse transcription (TPRT) mechanism. Based on the previous observations, following objectives for the research work were set-

1) To check the expression status of EhLINE1/EhSINE1 and identify the maximally expressed and silent copies.

2) Effect of cytosine DNA methylation in the promoter region on EhLINE1 expression.

3) Expression and enzymatic activity of EhLINE1–encoded ORF2 polypeptide.

Materials and Methods

2.1 Sources of materials and chemicals

E. coli strain DH5α was obtained from Bethesda Research Labs (BRL, USA) and was used for all recombinant DNA work. *E. coli* strain RIL was a kind gift from Prof. DN Rao, IISc, Bangalore, India. Plasmid vector pBluescript II KS+ was obtained from Stratagene (U. S. A.). Restriction enzymes and other molecular biological reagents were purchased from New England Biolabs (NEB, USA), Roche Biochemical's (Germany), Amersham Pharmacia (USA), Promega (USA), Sigma-Aldrich (USA), MBI Fermentas (Canada) and Qualigens (India). Random priming kit for labeling DNA was obtained from Thermo Fisher Scientific (USA); PCR purification kit was from Qualigens; Oligonucleotide primers were synthesized by Sigma (USA); EZ DNA Methylation-Lightning kit for DNA methylation was purchased from Zymo Research (USA); [γ-32P]-ATP and [α-32P]-dATP (specific activity ~5000 Ci/mmol) were obtained from Board of Radiation and Isotope Technology (BRIT, India). Adult Bovine Serum was purchased from PAA Laboratories (Austria) and *E. histolytica* media components were obtained from Sigma-Aldrich and DIFCO (USA). Diamond's vitamin mix for *Entamoeba* culture was purchased from Sigma-Aldrich (USA). X-Ray films were from Kodak and charged nylon membranes (GeneScreen plus) was obtained from New England Nuclear (NEN, USA).

(All concentrations indicated in percentage are (w/v) unless stated otherwise. All solutions were prepared in double distilled water unless stated otherwise. Autoclaving was done at a pressure of 15 lbs per square inch for 20min).

2.2 Organisms and growth conditions

E. coli DH5-α has the genotype: *SupE*44 *lac*U169 (ϕ80 *lacZ* M15) *hsdR17 recA1 endA1 gyrA*96 *thi*-1 *relA*1. Cells from an agar stab or frozen glycerol stock were first streaked on an LB plate (containing the appropriate antibiotic wherever necessary) and allowed to grow overnight at 37°C. Liquid cultures in LB medium were initiated from a single colony and were grown with constant shaking at 225 rpm at 37°C. Cells grown overnight was diluted 100 times which was further used as inoculum in fresh LB medium and grown with aeration at 37°C for 3-4h to obtain log phase cultures.

E. histolytica strain HM-1: IMSS clone 6 was a kind gift from Dr. William A. Petri (University of Virginia, USA); all experiments were done with *E. histolytica* strain HM-1: IMSS clone 6. The cells were maintained and grown in TYI-33 medium complemented with 15% adult bovine serum, 1X Diamond's vitamin mix and antibiotic (125µl of 250U/ml Benzyl Penicillin and 0.25mg/ml Streptomycin per 90ml of medium).

2.3 Culture media

2.3.1 LB Medium Bacterial cells were grown in Luria Broth (LB). It was prepared by dissolving 25gm of LB powder (Sigma-Aldrich) in 1 liter of distilled water and pH adjusted to 7.0 using 2N NaOH. The medium was sterilized by autoclaving.

2.3.2 LB Agar

LB agar was prepared by adding 1.5 % (w/v) of Bacto-Agar to LB medium and sterilized by autoclaving. Ampicillin 100µg/ml or Kanamycin 60µg/ml to a final concentration was added (when required) after cooling the LB agar to around 55ºC and plates poured under aseptic conditions.

2.3.3 TYI-S-33 medium composition per 900 ml (10 units) (Diamond et al. 1978)

Potassium phosphate, dibasic	1.0g	
Potassium phosphate, monobasic	0.6g	
Biosate peptone	30.0g	
Dextrose	10.0g	
Sodium chloride	2.0g	
L-Cysteine hydrochloride	1.0g	
Ascorbic acid	1.0g	
Ferric ammonium citrate	22.8	mg

The mentioned components were added in 700ml of double distilled water and pH adjusted to 6.8 using 2N NaOH. The volume was made up to 900ml and filtered using Whatmann #1 filter paper, aliquoted and autoclaved. Incomplete medium was stored at -20ºC. The medium was completed by adding 15% heat inactivated adult bovine serum, 1X Diamond's vitamin mix and 125µl of antibiotic (250U/ml Benzyl Penicillin and 0.25mg/ml Streptomycin). This media is used to grow *E. histolytica.* For heat stress, mid log phase *Entamoeba* cells grown in complete medium were treated at 42ºC for 1h.

2.4 Heat inactivation of serum

Adult bovine serum was stored at -20°C. Before heat inactivation, the serum was thawed at room temperature (RT) and incubated in a water bath at 37°C for 30min with intermittent shaking. The serum was then transferred to 55°C for 45min with intermittent shaking for complement inactivation and stored at 4°C.

2.5. Preparation of plasmid DNA from *E. coli* transformants

2.5.1 Mini-preparation of plasmid DNA (Alkaline lysis method) (Birnboim and Doly, 1979)

A single colony harboring the desired plasmid was inoculated in 2ml of LB medium containing appropriate antibiotic and grown overnight at 37^{o}C with shaking at 225 rpm. The cells were pelleted at 6000 rpm for 5min and the supernatant was aspirated out. The pellet was suspended in 100µl of Solution I (50mM Glucose, 25mM Tris-Cl pH 7.5 and 10mM EDTA pH 8.0). To the tube 200µl of freshly prepared Solution II (0.2N NaOH and 1% SDS) was added, mixed gently by inverting and incubated at RT for 5min, 150µl of chilled Solution III (3M Potassium acetate, pH 5.2) was then added and the contents were mixed gently by inverting the tube and kept on ice for 10min. The mixture was centrifuged at 13000 rpm for 10min at 4^{o}C. The supernatant was transferred to a fresh tube and 0.7 volumes of isopropanol was added and centrifuged at 13000 rpm for 10min. The pellet was washed with 70% ethanol by centrifugation at 13000 rpm for 5min at RT. The supernatant was discarded and the pellet was air-dried. The dried pellet was suspended in autoclaved Milli-Q or TE-RNAse.

2.5.2 Agarose gel electrophoresis

DNA fragments of size > 400bp were resolved on 0.8% agarose gel, while those in the range of 250 – 500bp were resolved on 1.2% agarose gel, 1.5% agarose was used for DNA fragments of size < 250bp. The gels were electrophoresed in 0.5 X TBE buffer containing 0.5µg/ml ethidium bromide.

2.5.3 Elution of DNA from agarose gel

The agarose slice containing the band of interest was cut out from the gel and chopped. DNA from gel band was isolated using QIAquick Gel Extraction Kit manufacturer's protocol (Qiagen). Eluted DNA checked by agarose gel electrophoresis and the concentration was measured by Nanodrop spectrophotometer.

2.6 DNA manipulations for cloning purposes

2.6.1 Polymerase Chain Reaction (PCR)

Forward and reverse oligonucleotide primers flanking the desired region of interest were used for PCR. The oligonucleotides used for various PCR reactions and their sequences are given in the appendix. All PCR reactions were performed by Taq DNA polymerase; phusion DNA polymerase and ZymoTaq DNA polymerase. A typical amplification reaction contained 50ng (plasmid) or

150ng (Genomic DNA) of template DNA, 1X Taq polymerase buffer, 200µM dNTPs, and 20 pmoles each of forward and reverse primers, 2.5mM MgCl2 and 0.5U of Taq or Phusion enzyme (for amplicons greater than 1kb) was used in a reaction volume of 50µl. For ZymoTaq DNA polymerase, 250µM dNTPs and 2U of enzyme along with 1x buffer used per 50ul reaction. The PCR cycle comprised of an initial denaturation at 94°C for 5min followed by 30 cycles of denaturation at 94°C for 30s, annealing at the Tm of the primers used for 1 min, extension at 72°C for 0.5-2min (1min / 1000bp for Taq polymerase, and 1min / 2000bp for Phusion). The last extension step at 72°C was done for an additional 10min. For ZymoTaq DNA polymerase Cycling conditions were: 95^{o}C/10min, 40 cycles of 95^{o}C/30s, 50-55^{o}C/30s, 72^{o}C/30-60s subsequently followed by 72^{o}C/7min. The amplification reaction was carried out in a DNA thermal cycler (Applied Biosystems, USA). The size and integrity of the products were checked by electrophoresing 5µl of the sample on a 0.8-1.2% agarose gel at 4 V/cm for an appropriate time period.

2.6.2 Restriction enzyme digestion of DNA

All restriction digestions were carried out according to the manufacturer's recommendations. The digestions were carried out in a water bath set at the recommended temperature. For analytical purpose, the reactions were set up in a volume of 20µl. For preparative purposes, the digestions were set up in a volume of 50 –100µl. After incubation, the reaction mixtures were loaded with 1X gel loading buffer (GLB) (2.5 % Ficoll type-400, 0.04% bromophenol blue, 0.04% xylene cyanol FF) onto an agarose gel.

2.6.3 Dephosphorylation of DNA termini

5'-phosphate groups from DNA fragments were removed using Antarctic Phosphatase (AP) from NEB. The digested vector DNA (1µg/µl) was resuspended in 1X AP reaction buffer and incubated with AP at 37^{o}C for 20 to 30mins. The enzyme was heat inactivated at 65^{o}C for 5min and DNA was purified by gel extraction.

2.6.4 Ligation of DNA termini

Ligation **r**eactions were carried out in a volume of 10µl at 16°C for 16h. The reaction mixture contained about 50ng of the digested vector DNA, insert DNA fragment at 1:3 or 1:5 (vector: insert) molar concentrations and 1X ligase buffer containing 1mM ATP and 20U of T4 DNA ligase. The ligation reaction product was transformed into *E. coli* (DH5α) competent cells and transformants were selected on LB agar plates supplemented with the required antibiotic.

2.6.5 Preparation of competent cells

Competent cells of *E. coli* (DH5α, BL21 (RIL) and Rosetta) were prepared by the method described by Hanahan *et al*. 1991. A single colony from LB agar plate was inoculated into 5ml LB medium and grown overnight at 37ºC. 1% of the overnight culture was added to 50ml LB and grown at 30ºC to an OD600 of 0.38-0.42. The culture was then centrifuged at 5000 rpm for 10min at 4ºC. The supernatant was decanted and the pellet was suspended gently in 10ml of prechilled 0.1 M CaCl2. The cells were incubated for 1h on ice and then centrifuged at 5000 rpm for 10min at 4ºC. The supernatant was discarded and the pellet was re-suspended in 2ml of 0.1M CaCl2. The competent cells were stored as15% glycerol stocks in 100µl aliquots at -80°C.

2.6.6 Transformation of competent cells

100µl competent cells were thawed on ice and 5-10ng of plasmid DNA was added. The cells were incubated on ice for 30mins. Cells were then given heat shock at 42°C for 90s and incubated on ice for 5mins. 1.9ml of LB was added to the cells and the cells were grown at 37ºC for 1-1.5h at 200 rpm. Cells were then pelleted down at 5000 rpm for 10mins at 4°C, cells pellet was resuspended in 100µl LB and plated on LB agar plates with appropriate antibiotic followed by incubation at 37°C for 14-16h. Transformants were further screened by plasmid isolation and restriction digestion.

2.7 Isolation of total RNA from *E. histolytica* trophozoites

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

2.7.1 Isolation of poly(A)+ RNA

Poly(A)$^+$ RNA was purified from total RNA obtained from *E. histolytica* trophozoites using poly(A) tract mRNA isolation system III from Promega (Z5300) as per manufacturer protocol. The System uses Magnesphere technology to isolate mRNA effectively from total RNA with the maximum capacity of 1mg total RNA per column. Briefly, total RNA was incubated in a sterile tube at 65ºC/10min. A biotinylated oligo(dT) probe and SSC was added to the RNA and incubated at room temperature for 10min to hybridize with the 3'-poly(A)$^+$ region in mature mRNAs. 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 highly pure mRNA was eluted from the solid phase by the addition of provided ribonuclease-free, deionized water.

2.7.2 Analysis of RNA

RNA samples were run on 1.2% denaturing agarose gels containing 2.2M formaldehyde prepared in 1X MOPS buffer (20mM MOPS, 2mM sodium acetate and 1mM EDTA). Glassware used for RNA isolation and analysis were treated with 0.1% (v/v) DEPC (Diethyl pyrocarbonate solution in water) for 10-16h at 37ºC followed by baking at 180ºC for 8h as described by Sambrook *et al*. 1989. Electrophoresis chamber used for RNA samples was treated with 3% (w/v) hydrogen peroxide solution for 30min in dark and washed extensively with DEPC treated water.

2.7.3 Diethyl pyrocarbonate (DEPC) treatment of reagents

0.1 % DEPC was added to the solution to be treated and shaken vigorously to bring the DEPC into the solution and incubated for 12h at 37ºC. After incubation, solutions were autoclaved to remove traces of DEPC.

2.7.4 Northern blotting

RNA (20-30µg) was denatured by incubating with 2X RNA loading dye at 65ºC for 15min followed by chilling on ice. Samples were centrifuged briefly and mixed with 2X RNA loading dye (Fermentas) and then loaded on to denaturing agarose gel containing 2.2M formaldehyde and 1X MOPS buffer. Electrophoresis was carried out at 4 V/cm. The gel was then washed extensively with DEPC treated water to remove the formaldehyde. The gel was then sequentially treated for 20min each with a denaturing (0.05N NaOH and 1.5M NaCl) and neutralizing solution (0.5M Tris-HCl pH 7.5 and 1.5M NaCl) followed by 20min equilibration in 20X SSC. The transfer membrane was pre-equilibrated with 20X SSC and the transfer was carried out using the 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.

2.8 Hybridization of radiolabeled probes to immobilized nucleic acids

2.8.1 Preparation of radiolabeled DNA by random priming method

About 50-100ng of linear DNA along with Decanucleotide in 5X Reaction Buffer was denatured by heating in a boiling water-bath for 10min and immediately chilled on ice. To the tube containing denatured DNA; 3µl of mixA, 30-50µCi [α-32P] dATP and 5U of Klenow enzyme added. The reaction was initiated by incubating at 37ºC for 5min followed by repeated incubation with 4ul of dNTP mix at 37ºC for 5min. After incubation 1µL of 0.5M EDTA, pH 8.0 added to stop the reaction. All the components used were from Thermo Scientific; DecaLabel DNA Labeling Kit (#K0622). Unincorporated dNTPs were removed by ethanol precipitation in the presence of 50µg of carrier DNA (salmon sperm DNA) and 2.5M ammonium acetate, or by nucleotide removal kit (Qiagen).

2.8.2 Generation of radiolabeled strand specific probe

In order to synthesize 32p-labeled single stranded DNA probes, ORF1 full-length (14-1506nt) and ORF2 B+C region (2470-3615nt) was used as a template after amplification with the primer set HJ67 FP+EK39 RP and BK49 FP+DY32 RP respectively. Purified template together with one primer from each primer set used for the linear PCR (LPCR). The reaction protocol used as mentioned by Millican *et al* (1997) with few modifications. Amplification reaction contained 1X Taq polymerase buffer, 200µM each dA/G/TTP, and 5µM dCTP, 50µCi [α-P^{32}] dCTP, 30pmol respective primers, 10ng/kb DNA template and 5U of *Taq* DNA polymerase in a reaction volume of 50µl. The linear PCR cycle comprised of an initial denaturation at 94°C for 3min followed by 40 cycles of denaturation at 94°C for the 30s, annealing at the Tm of the primer used for 1min, extension at 72°C for 1min 30s. The last extension step at 72°C was done for an additional 10min. Unincorporated dNTPs were removed by nucleotide removal kit (Qiagen).

2.8.3 Hybridization

RNA blots were first incubated in prehybridization solution (1% SDS and 1M NaCl, 0.3-0.4ml per square cm of the membrane) at 65ºC in hybridization bottles. After 3h, heat-denatured radiolabeled probe (2 x 105 dpm/ml) and 100µg/ml denatured salmon sperm DNA was added to the prehybridization mix and hybridization was carried out for 16h at 65ºC. The membranes were

washed sequentially to remove non-specifically bound probe using the following protocol: twice with 2X SSC at RT for 5min, twice with 1X SSC and 1% SDS at 65ºC for 30min and finally twice with 0.1X SSC at RT for 30min each. After washing membrane was covered with saran wrap and autoradiographed.

2.8.4 Removal of probe from nylon membrane for rehybridization

The probe was stripped off the membrane by incubating the membrane in a boiling solution of 0.1X SSC and 0.1 % SDS for 20-30min. The efficiency of probe removal was monitored by exposing the blot to a phosphorimaging film before hybridization.

2.8.5 Autoradiography

After hybridization and washing, the blots were wrapped in saran wrap and mounted. Autoradiography was performed with phosphorimaging film (GE Healthcare) for the appropriate amount of time.

2.9 DNAse I digestion of *E. histolytica* RNA

15µg of total RNA was taken in a microfuge tube and to it, 3µl of 10X DNase I reaction buffer (40mM Tris-HCl, 2mM MgCl2) and 1µl of DNase I enzyme (10U, MBI Fermentas) and RNase free water were added up to 30µl. The tube was incubated at 37ºC for 45min. and DNase I was inactivated by incubating at 65ºC for 15min in presence of 10mM EDTA, pH 8.0.

2.10 Reverse transcription PCR (RT-PCR) assay

3-5µg of total RNA (after DNase I treatment) was taken in a microfuge tube and to this 100pmole of oligodT or gene specific primer and 1.0mM dNTP was added and volume was made up to 12µl. The mixture was incubated at 65ºC for 5min followed by quick chill on ice. To it 4µl of 5X first strand buffer (50mM Tris-HCl pH 8.3, 75mM KCl and 3mM MgCl2), 0.5µl of RNasin plus RNase inhibitor (40U/µl, Promega), 1µl of 0.1M DTT and 1µl of Superscript III reverse transcriptase (200 U/µl, USB) was added. The reaction mixture was incubated at 50ºC for 1h and reverse transcriptase was inactivated by incubation at 70ºC for 10min. 2-5µl of this mix was used for a regular PCR reaction with gene specific primers.

2.11 Real Time PCR

2.11.1 Primer design

Real-time PCR (qRT-PCR) primers were designed taking two factors into account:

(1) primer should specific and should end in G/C

(2) Tm of the primer should be high 58-64^{0}C.

Each primer was analyzed for homology with the *E. histolytica* database and any primer that had significant sequence similarity to multiple genes was rejected. Thus, both the forward and reverse primers were specific for one gene. Optimal annealing conditions were used to ensure specificity and any PCR primer pair that produced more than one melt peak was discarded.

2.11.2 Quantitative Real Time (qRT-PCR)

DNase I treated total RNA (400ng) was reverse transcribed using qHspRP into cDNA by Revertaid-RT (Fermentas). Real-time quantitative PCR was performed in 7500 Real Time PCR System (Applied Biosystems) using SYBR green PCR Master Mix, 2pmol of qHspFP, qHspRP and 2µl of cDNA (serial 1:10 fold dilution). Actin (control gene) were amplified in parallel. The conditions were pre-denaturation at 95°C for 10min, followed by 40 cycles at 95°C for 15s and 58°C for 1min followed by a dissociation stage at 95°C for 15s and 58°C for 1min. Cycle threshold values (C_t) were analyzed by the SDS1.4 software (Applied Biosystems) and all samples were analyzed in triplicates in three independent experiments. Reactions without cDNA were used as no template control and no RT controls were also set up to rule out genomic DNA contamination.

2.12 DNA substrate preparation for Endonuclease assay

2.12.1 Supercoiled plasmid DNA preparation

Supercoiled pBS plasmid was purified using Qiagen plasmid purification kit and DNA concentration was estimated by measuring absorption at 260nm.

2.12.2 Endonuclease assay with pBS supercoiled DNA

Cleavage reactions were carried out in a buffer containing 50mM Tris-Cl (pH 7.5), 100mM NaCl, 10mM $MgCl_2$ and 1mM DTT at 37^{o}C as used previously for Eh EN (Yadav *et al*., 2010). Enzyme and substrate concentrations were used as indicated in each experiment. Reactions were stopped by removing aliquots of 20µl and mixing them with 5µl of stop mix (100mM EDTA, 10mM Tris-Cl (pH 8.0), 30% glycerol, and 0.25% bromophenol blue). Each sample was then analyzed by electrophoresis through 0.8% agarose in Tris-borate (45mM Tris-borate, 1mM EDTA) containing 0.5µg/ml ethidium bromide at 3V/cm. Under these conditions, the covalently closed circular form of pBS migrated fastest, followed by linear and open circle forms.

2.13 Isolation of genomic DNA from *Entamoeba* trophozoites

Entamoeba cells (approximately 4 x 10^7 cells) were harvested by chilling on ice for 10min and centrifuged at 280g at 4°C for 7min and washed once with PBS # 8 [0.37% K_2HPO_4, 0.11% KH_2PO_4 and 0.95% NaCl, pH 7.2]. The cell pellet was resuspended in 2ml of nuclei lysis solution (Promega) and was pipetted to lyse the cells. 75µg of RNase A was added, mixed by inverting and incubated at 37°C for 30min. To this, 1.0ml of protein precipitation (Promega) solution was added, vortexed and was kept on ice for 5min. The suspension was centrifuged at 16000g at 4°C and supernatant was transferred to a new tube. The DNA was precipitated by addition of isopropanol and was centrifuged at 16000g for 10min at room temperature. The pellet was washed with 70% ethanol and was dried and dissolved in 100µl of TE.

2.14 Bisulfite treatment of Genomic DNA, PCR amplification, and cloning

Bisulfite (BS) modification of genomic DNA was carried out with the EZ l Methylation-Lightning Kit (Zymo Research) following the manufacturer's protocol. Briefly, lightning conversion reagent (provided in the kit) was added directly to the DNA (500ng) to be treated and the reaction was performed in a thermal cycler for 98ºC/8min followed by 54ºC/60min. Desulfonation and clean-up of the bisulfate converted DNA was performed using the Zymo spin column (provided in the kit). Bisulfite treated DNA was PCR amplified in a 50µl reaction containing 0.25mM each dNTP, 2U zymoTaq DNA polymerase along with 1µM BS-converted (or normal) primers/nested primers. Cycling conditions were: 95ºC/10min, 40 cycles of 95ºC/30s, 50–55ºC/30s, 72ºC/30–60s subsequently followed by 72ºC/7min. Amplified products were either sequenced directly or cloned into pGEMT-EASY vector (Promega) and subjected to Sanger sequencing.

2.15 Single nucleotide incorporation

Bisulfite treated/untreated DNA (200ng) was mixed with 50µM dGTP/dATP, phusion buffer (1×), labeled primer (~60,000 counts) and 1U phusion polymerase (NEB). After heating at 95ºC/5min, annealing was done at 50ºC/2min, followed by incorporation at 72ºC/10min. The product was denatured at 95ºC/5min followed by snap chill on ice/5min and separated on 7M- 6% Urea-PAGE. Incorporation was detected by Phosphor Imager (Typhoon FLA 9500, GE).

2.16 End-labeling of synthetic oligo

100pmol of oligo was taken in a microfuge tube and to it polynucleotide kinase (PNK) buffer was added to a final 1X concentration (70mM Tris-HCl (pH 7.6), 10mM MgCl2, 5mM DTT)), 40µCi of [γ-32P] dATP and volume made up to 24µl by nuclease free water. The reaction was initiated by

addition of 1µl of T4 polynucleotide kinase (10U/µl, NEB). The reaction mix was incubated at 37ºC for 1h. The reaction was stopped by incubating at 65ºC for 20min. Labeled oligos were purified by Nucleotide removal kit (Qiagen).

2.17 Primer Extension

DNase I treated total RNA (10µg) was incubated with [γ-32P] ATP end labeled primer. Annealing was carried out at 65ºC for 5min, followed by extension at 42ºC for 1h with 200U Superscript III reverse transcriptase (Invitrogen). The products were separated on denaturing 6% urea-polyacrylamide gel, together with the sequencing reaction using the same oligonucleotide.

2.18 Denaturing polyacrylamide gel electrophoresis

The glass plates used Owl apparatus) were cleaned thoroughly with detergent. The sandwich was assembled as per manufacturer's instruction. 30ml of the gel mix was prepared containing 6-15% Acrylamide (Acrylamide: bisacrylamide=19:1) and 7M urea in 1X TBE. 200µl of freshly prepared 10% ammonium persulfate (APS) and 20µl TEMED. The contents were mixed and poured immediately into the sandwich taking care that no air bubble was introduced. The gel was used 1-2h after pouring. The pre-run was performed in 1X TBE for 30-45min. Samples were denatured by heating at 95ºC for 4-5min and snap chilled just before loading. Electrophoresis was carried out at 150V for 2-3h. After the electrophoresis was over, the sandwich was dismounted and disassembled. The gel which was attached to the notched plate was transferred to Whatman #3 and dried in a gel dryer at 80ºC for 2h. Dried gel was exposed to an imaging plate (IP) and scanned in a phosphorimager.

2.19 DNA sequencing

The glass plates used for sequencing gels were thoroughly cleaned with detergent and treated with sigmacote (siliconizing reagent). The sandwich was assembled as per manufacturer's instruction. 60ml of gel mix was prepared containing 6% acrylamide (acrylamide: bisacrylamide = 19:1), 7M urea in 1X TBE. The solution was filtered through Whatman No 1 filter paper and to it 400ml of freshly prepared 10% ammonium persulphate and 55ml of TEMED was added. The contents were mixed properly and poured immediately into the sandwich taking care that no air bubble was introduced. The gel was kept for 2-4h to set. The temperature of the gel was brought up to 50ºC by performing a pre-run at 40W in 1X TBE for 30-45min. Samples were denatured by incubating at 75ºC for 5min. just before loading. The gel was run at 50ºC at 40-50W. Electrophoresis was terminated when bromophenol blue dye front reached the bottom of the gel. After run was over, the

gel was directly transferred to a Whatman 3mm sheet, covered with saran wrap and dried using gel dryer at 80^{o}C for 1h. Dried gel was exposed to imaging plate for requisite period and scanned in phosphorimager (Typhoon FLA 5000).

2.20 Transfection of *E. histolytica* trophozoites by electroporation

Transfection was performed by electroporation as described by Sahoo *et al*. 2004. Briefly, trophozoites in log phase were harvested and washed with PBS followed by incomplete cytomix buffer (10mM K_2HPO_4/KH_2PO_4 (pH 7.6), 120mM KCl, 0.15mM $CaCl_2$, 25mM HEPES (pH 7.4), 2mM EGTA, 5mM $MgCl_2$). The washed cells were then re-suspended in 0.8ml of complete cytomix buffer (incomplete cytomix containing 4mM adenosine triphosphate, 10mM glutathione) containing 200µg of plasmid DNA of construct to be transfected and subjected to two consecutive pulses of 3000V/cm (1.2kV) at 25µF (Bio-Rad, electroporator). The transfectants were initially allowed to grow without any selection. Drug selection was initiated after 2 days of transfection in the presence of 10µg/ml G418 for constructs with luciferase reporter gene.

2.21 Luciferase reporter constructs [P-ORF1 and P-ORF2]

Forward and reverse primers P-ORF1FP, P-ORF1RP, P-ORF2FP, and P-ORF2RP (sequences mentioned in the appendix) were used for amplification of the desired region. Parental cloned full-length EhLINE1 plasmid used as a template along with 1X Taq polymerase buffer, 200µM dNTPs, 20pmoles each of forward and reverse primers and 2U of Taq polymerase in a reaction volume of 50µl. Amplified PCR products were purified using Gel extraction kit (Qiagen) and cloned upstream of luciferase reporter gene in place of Lectin promoter in vector pEh-NEO-LUC Actin in which the 3'-end of Actin was cloned downstream of LUC. Constructs were transfected by electroporation and maintained in presence of G418 (10µg/ml) till the cells become stable.

2.22 Luciferase reporter Assay

The procedure was done as described by Shiteshu *et al*. Briefly, stably transfected trophozoites, were washed in PBS (pH 7.4), lysed in 200µl of reporter lysis buffer (Promega) with the addition of protease inhibitor cocktail (Sigma) and were frozen overnight at −80°C. Thawed on ice and pelleted to remove cellular debris. Before measuring the activity, Samples were allowed to warm at room temperature, measured according to the manufacturer's instructions (Promega) using a Luminometer (Promega) and activity per µg of protein calculated.

2.23 Total cell lysate preparation

One million trophozoites growing in log phase were harvested at 600 g for 5min at 4°C. The pellet was washed with cold PBS # 8 and then resuspended in 10mM Tris-Cl pH 7.5, 150mM NaCl, 1% Triton- X100, 2mM PHMB and 1X protease inhibitor cocktail (Sigma). The lysate was freeze thawed thrice and sonicated for 10s to shear the genomic DNA and centrifuged at 13000g for 5min to pellet down the debris. The supernatant was aliquoted and stored at -80ºC and quantification was done by BCA.

2.24 Expression and purification of recombinant proteins

E. coli expression strains (BL21, BL21-RIL, Rosetta etc) were transformed with desired plasmid (pET 30b, pET21a or pGEX4T-1) containing the DNA sequence for the protein to be expressed. Transformed cells were inoculated in LB medium containing appropriate antibiotic at 37ºC with aeration for 12 to 14h with shaking at 220 rpm. 2% secondary inoculum was given in 1L LB/TB medium and grown further at 37ºC to an ODA600 of 0.5-0.7. For induction of 6XHis or GST tagged proteins, IPTG was added to a final concentration of 0.5mM and the cells were further allowed to grow at 18ºC for 6-9h.

2.24.1 Purification of His tagged protein

Recombinant proteins were purified using Ni-NTA agarose affinity chromatography as described by the manufacturer's protocol. Cell pellet was resuspended in 40ml lysis buffer (50mM Tris-Cl pH 8.0, 300mM NaCl, 0.5-1 % Triton X-100, 10mM βME, 5- 20mM imidazole, 0.5mg/ml lysozyme, 1mM PMSF and 10% glycerol). The cell suspension was further incubated on ice for 30min followed by 10 cycles of sonication (30s on/1min off). The lysate was centrifuged at 12,000g for 30min at 4ºC. The supernatant was incubated with 1ml of pre-equilibrated Ni-NTA agarose (Qiagen) for 2h at 4ºC with gentle mixing. It was then packed in a C10/10 column (Amersham) and washed several times with wash buffer (50mM Tris-Cl pH 8.0, 300mM NaCl, 2mM βME and 10% glycerol) containing 20 to 100mM imidazole. Bound protein was further eluted with 200mM imidazole containing wash buffer. Fractions containing purified protein were identified by SDS-PAGE and then pooled and dialyzed against dialysis buffer C [50mM Tris-Cl (pH 7.5), 100mM NaCl, 2mM β-mercaptoethanol and 30% glycerol]. The purified protein was quantified by Bradford's method and stored in aliquots at –80°C.

2.24.2 Purification of GST tagged protein

Recombinant proteins were purified using glutathione sepharose 4 fast flow (GE Healthcare) affinity chromatography. Cell pellet was resuspended in 40ml lysis buffer [1xPBS (140mM NaCl, 2.7mM KCl, 10mM Na2HPO4, 1.8mM KH2PO4) pH 7.3, 5mM DTT, 0.5-1 % Triton X- 100, 5-20mM imidazole, 0.5mg/ml lysozyme, 1mM PMSF and 10% glycerol). Cell suspension was further incubated on ice for 30min followed by 10 cycles of sonication (30s on/1min off). The lysate was centrifuged at 12,000g for 30min at 4ºC. The supernatant was incubated with 0.25ml of pre-equilibrated glutathione sepharose 4 fast flow (GE Healthcare) for 4h at 4ºC with gentle mixing. It was then packed in a C10/10 column (Amersham) and washed several times with wash buffer (1X PBS, 3mM DTT and 10% glycerol). Bound protein was further eluted with elution buffer (50mM Tris-Cl pH8.0, 20mM glutathione reduced, 5mM DTT). Fractions containing purified protein were identified by SDS-PAGE and then pooled and dialyzed against dialysis buffer [50mM Tris-Cl (pH 8), 100mM NaCl, 5mM DTT and 30% glycerol]. The purified protein was quantified by Bradford's method and stored in aliquots at –80°C.

2.25 Protein estimation

2.25.1 BCA assay

The amount of protein in a sample was estimated by the bicinchoninic acid (BCA) assay using BSA as the standard. The working solution was prepared by mixing BCA (Sigma) and 4% copper sulfate in a ratio of 50:1. Equal volumes of the sample and the working solution were mixed in a microtiter plate and incubated at 37ºC till a purple color develops in the lowest concentration of BSA (~ 30 min). The absorbance was taken at 560nm using a microtiter plate reader (Thermo Scientific, USA).

2.25.2 Bradford's assay

The amount of protein in a sample was estimated by Bradford's assay using BSA as the standard. Bradford's reagent was prepared by dissolving 100mg of Coomassie Brilliant BlueG250 in 50ml methanol and 100ml H3PO4 (85%). It was diluted to 1000ml with MillQ water and filtered through Whatman no 2 and stored in a dark bottle at 4ºC. To quantify the protein, 100µl of protein sample (diluted) was mixed with 900µl of Bradford's reagent and absorbance (OD595) was measured after 5min in a spectrophotometer.

2.26 SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE was carried out under reducing conditions. The polyacrylamide gel was prepared using acrylamide (acrylamide:bis-acrylamide; 29:1) in 1.5% Tris-Cl pH 8.8, 0.1% (w/v) SDS, 0.04%

(w/v) APS and TEMED. After polymerization of resolving gel, stacking gel was poured. The stacking gel contained 4% acrylamide in 0.5% Tris-Cl pH 6.8, 0.1% (w/v) SDS, 0.04% (w/v) APS and TEMED. Samples were mixed with 4X SDS-PAGE loading dye to a final dye concentration of 1X. After electrophoresis, proteins were fixed in the gel by incubating in fixing solution (50% methanol, 7.5% acetic acid) and detected by Coomassie Brilliant Blue (0.25% CBB R-250 in fixing solution). The gels were destained in the fixing solution and dried.

2.27 Transfer of proteins (Western blotting)

Polyacrylamide gel to be transferred was incubated in Towbin buffer (For 500ml: 1.51g Tris base, 7.2g Glycine, 100ml Methanol, pH 8.3) for 15min. The treated gel was placed on two sheets of Whatman 3mm paper cut to the size of the gel, saturated with Towbin buffer. A sheet of PVDF membrane pre-activated by soaking in methanol followed by in Towbin buffer was placed on the gel taking care that no air bubble(s) were trapped in between the membrane and the gel. Two sheets of Whatman 3 mm paper were placed above the membrane. The transfer was set at constant milliamp (mA), depending on the size of the membrane (0.8 times the area of the membrane) for 1-1.5h. The membrane was then stained with Ponceau S and was blocked overnight at 4ºC with 5 % skimmed milk powder in PBS-T (PBS containing 0.05 % Tween 20). Primary antibody followed by secondary antibody incubation was done in 3 % milk powder in PBS-T with shaking at RT for 2h and 1h, respectively. The blots were washed thoroughly with PBS-T after every incubation with antibody. The secondary antibody used was horse radish peroxidase conjugated IgG. Band detection was done using ECL kit (Millipore). Antibody dilutions used: 1:1000, Anti-ORF1 (polyclonal, rabbit); 1:8000, Anti-GST antibody; 1:5000 Anti-His antibody; 1: 10000, Anti-Rabbit-HRPO; 1: 10000, Anti-Mouse-HRPO.

2.28 *In vitro* synthesis of RNA (Ribo Max large Scale RNA Production System-T7)

Linear DNA templates with desired end points were PCR amplified from genomic DNA of *E. histolytica*. T7 promoter sequence was incorporated in the forward primer. The amplified DNA fragment was gel purified under RNase free conditions. Before the start of reaction equal volume of four individual rNTPs (10mM of each ATP, GTP, CTP, and UTP) were mixed to produce a final solution containing 0.4mM of each nucleotide. In vitro transcription reaction was set up at room temperature (DNA could precipitate at a low temperature in the presence of spermidine, a component of transcription buffer). Reaction components were added in the following order.

T7 Transcription 5X buffer: 5µl

rNTPs (10mM each): 4µl

Linear DNA template (500ng): - µl,

Enzyme Mix: 2µl

Final Volume: 25µl

The reaction was gently mixed and incubated at 37ºC for 1h. After completion of the reaction template DNA was digested by incubating the transcription product with 5U DNase I (RNase-free) for 15min at 37ºC. Unincorporated nucleotides were removed by DNA precipitation using 1/10 volume of 3M *sodium acetate*, pH 5.2. Purified RNA was quantified and stored in –80°C till further use. For the synthesis of P32 labeled RNA following reaction has been set up:

T7 Transcription 5X buffer: 5µl
rATPs (10mM): 1µl
rGTPs (10mM): 1µl
rCTPs (10mM): 1µl
rUTPs (0.1mM): 2.4µl
αP32UTPs: 3 µl
Linear DNA template (500 ng): - µl
Enzyme Mix: 2µl
Final Volume: 25µl

The reaction was gently mixed and incubated at 37ºC for 1h and further processed as described earlier.

2.29 Dot Blot Assay:

The DNA was denatured by addition of NaOH to a final concentration of 0.25N, in a total volume of 200µl. After keeping the DNA at room temperature for 30min, it was transferred on to the ice. Whatman no III and Nylon membrane cut to the required size was saturated in 0.4M Tris-Cl pH 7.5 for 15min and assembled on to the dot blot apparatus. Denatured DNA samples were loaded to the membrane through dot blot wells. The blot was dried under vacuum pressure and used for probe hybridization after UV cross-linking.

2.30 Densitometric estimation

Band intensities were compared by using the AlphaEase software which provides the desired sensitivity.

2.31 Targeted sequencing and RNA sequencing (Illumina) analysis

2.31.1 Targeted sequencing

Targeted sequencing is done in IGIB, Delhi with the kind help of Dr. Kausik Chakraborty and Manish Rai. It is based on sequencing-by-synthesis and fusion PCR method using Ion torrent

platform which is similar to other platforms. It differs in that it does not uses fluorescence or chemiluminescence, instead, it works on the principle of detection of hydrogen ion released during the incorporation of nucleotides into the growing DNA template. Normally a hydrogen ion releases as a by-product when nucleotide incorporates to a DNA strand by the polymerase. The released H^+ from the reaction causes a change in pH, which is measured by the sequencer.

2.31.1.1 Bidirectional sequencing using the fusion primer

For the targeted sequencing, fusion primers (adapter sequence which was used in sequencing and primer for amplification of the insert) were used to amplify the amplicons from cDNA template. cDNA was synthesized with the respective reverse primer (without adapter sequence) from each amplicon primer set and further used as a template for the generation of amplicon with adapter containing primers. Primer sequences with and without adapter have been mentioned in the appendix section. The fusion primers were designed in such a way that it contains primer sequence for amplification of amplicons as well as adapter sequence, which will help in sequencing. Four fusion primers were used to prepare the amplicon library during fusion PCR including two pairs of forward and reverse primers per target region to enable bidirectional sequencing. Details about the ion torrent library preparation are available in ion torrent library preparation manual (Ion Amplicon Library Preparation (Fusion Method) Publication Number 4468326). The template was prepared from amplified PCR using ion one touch instrument and enrichment of positive ion sphere particle was done using Ion one touch ES instrument (Ion OneTouchTM Template Kit**,** 4468660). The prepared template was sequenced in ion torrent sequencer using the manual Ion Sequencing Kit User Guide v2.0, 4468997.

2.31.2 RNA sequencing with Illumina platform

Total RNA was extracted from the trophozoites using the Trizol method. The amount and integrity of the extracted RNA were determined by NanoDrop 2000c spectrophotometer (Thermo Scientific) and visually after electrophoresis on a 1.2% agarose gel containing ethidium bromide. Total three biological samples of RNA were provided to SciGenome for the construction of cDNA libraries with poly$(A)^+$ selection and sequenced on Illumina HiSeq 2500 by paired end deep sequencing, producing 2 × 100-nucleotide paired-end reads. In total, about 40 million reads were sequenced. Total reads were trimmed of low-quality and adapter sequences using trimmomatic-0.36 (Bolger *et al*., 2014) and quality checking was done by FastQC. Paired reads were allowed to map to the nucleotide sequence of total LINE1 copies (N = 967) of *Entamoeba histolytica* using RSEM (Li & Dewey, 2011).

2.31.2.1 Normalization of Gene Expression Levels and Identification of Differentially Expressed LINE1.

Sequencing reads were mapped to the reference sequences by RSEM, the expression level was measured by transcript per kilobase million (TPM) to make it easier to compare the proportion of reads that mapped to LINE1 in each biological sample. Differentially expressed LINE1 were looked by EBseq package (Leng and Kendziorski, 2015).

2.32 Bioinformatics tools

- All sequences were extracted from NCBI database or Amoeba DB (http://amoebadb.org/amoeba/).
- Protein sequence analysis was performed using Expasy tool (http://web.expasy.org).
- Bioedit sequence alignment editor was used for sequence alignment (http://www.mbio.ncsu.edu/bioedit/bioedit.html)

Results

3.1 Expression analysis of EhLINE1 and EhSINE1

Since expression of transposable elements could be potentially mutagenic, their transcription is tightly regulated in most organisms. The expression status of the multiple copies of EhLINEs and EhSINEs is not known. The *E. histolytica* genome has 967 copies of EhLINE1. Of these, only 57 copies are full-length (i.e.~ 4.8kb), while the rest is truncated at either the 5' or 3'-end, or at both ends. For EhSINE1 there are 493 copies of which 393 copies are full-length. The latter range from 451- 684bp in length due to varying numbers of internal repeats (Huntley *et al*., 2010). They have been categorized into seven classes on the basis of short tandem repeats of 26-27bp which are present in zero to four copies per element. Briefly, 'None' class has no repeat; 'R1' class copies have one repeat of R1 type; 'R2' has two repeats of R1 and R2 type; 'R3' has three repeats of R1, R2, and R3 type; 'R4' have four repeats of R1, R2, R3 and R4 type. In addition, there are some copies with a size corresponding to three repeats of which two are degenerate and only one is recognizable. The 'R1 only' class has the recognizable repeat of R1 type, while 'R3 only' contains a recognizable repeat of R3 type. The present study was undertaken to determine the expression status of the multiple copies of EhLINE1and EhSINE1.

3.2 Distribution of EhLINE1 and EhSINE1 on the basis of size

To draw a correlation between the expression status of EhLINE1/SINE1 with the type of the copy, we categorized all EhLINE1 and EhSINE1 copies on the basis of their size. We extracted all the EhLINE1/EhSINE1 copies from the database and aligned them using MAFFT, a multiple sequence alignment tool, using default parameters. In EhLINE1, copies were assigned as 5'-intact, 3'-intact and both side truncated on the basis of their alignment to the full-length reference sequence. Finally, 57/967 copies of EhLINE1 were assigned as full length and rest were truncated. In truncated, 199 copies showed 3'-truncation and 5'-intact; 330 copies were 5'-truncated and 3'-intact and 381 copies were assigned as truncated from both the side as shown in Figure 12. Similarly, for SINE1 393 copies were assigned as full length with different repeat sequences and additionally, 100 copies were found to be truncated.

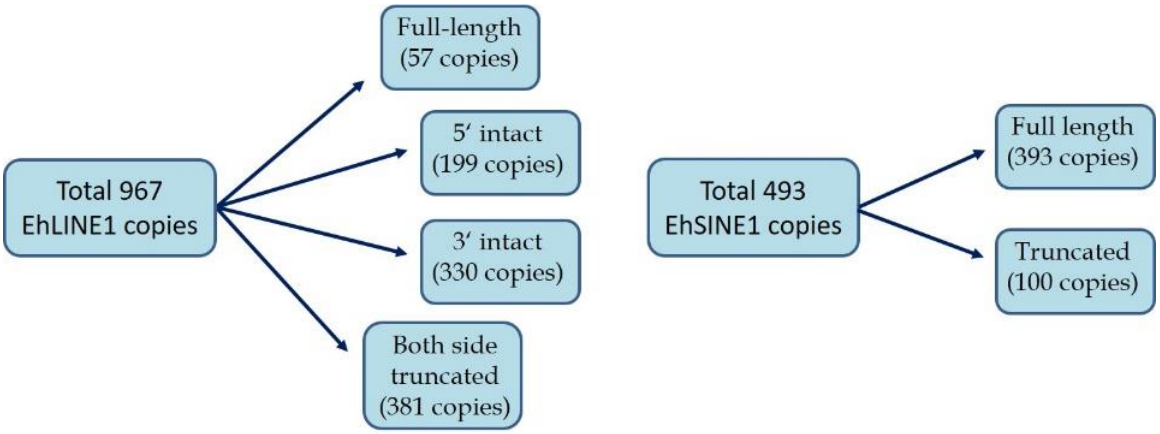


Figure 12: Schematic representation of EhLINE1/SINE1 copies derived from sequence alignment

3.3 Expression status of EhLINE1 and EhSINE1 by targeted sequencing of expressed transcripts

To identify the copies that might be transcriptionally more active than others we decided to do an expression analysis of EhLINE1/EhSINE1 through targeted amplicon sequencing by using amplicons from cDNAs of these elements, and sequencing them on Ion Torrent platform, as described in Methods section. The rationale was to design PCR primers from conserved regions of these elements which flank variable regions. Sequencing of these amplicons should identify the expressed copies based on sequence match. To select the regions suitable for designing primers we aligned all the full-length EhLINE1 and EhSINE1copies and looked for the highly similar regions. Six such regions were selected from the 4.8kb EhLINE1 to cover the entire element and primers were designed from these (Fig.13). For EhSINE1 we selected only one region due to its small size, and one primer set was designed from the maximally conserved region as shown in Figure 13. Barcode sequence with two different adapter sequences in forward and reverse primers were added at the 5'-end of oligos to enable sequencing from both ends. These barcode and adapter sequences together increased the amplicon size by 43 nucleotides. The size of each amplicon with and without an adapter is given in Table 3.

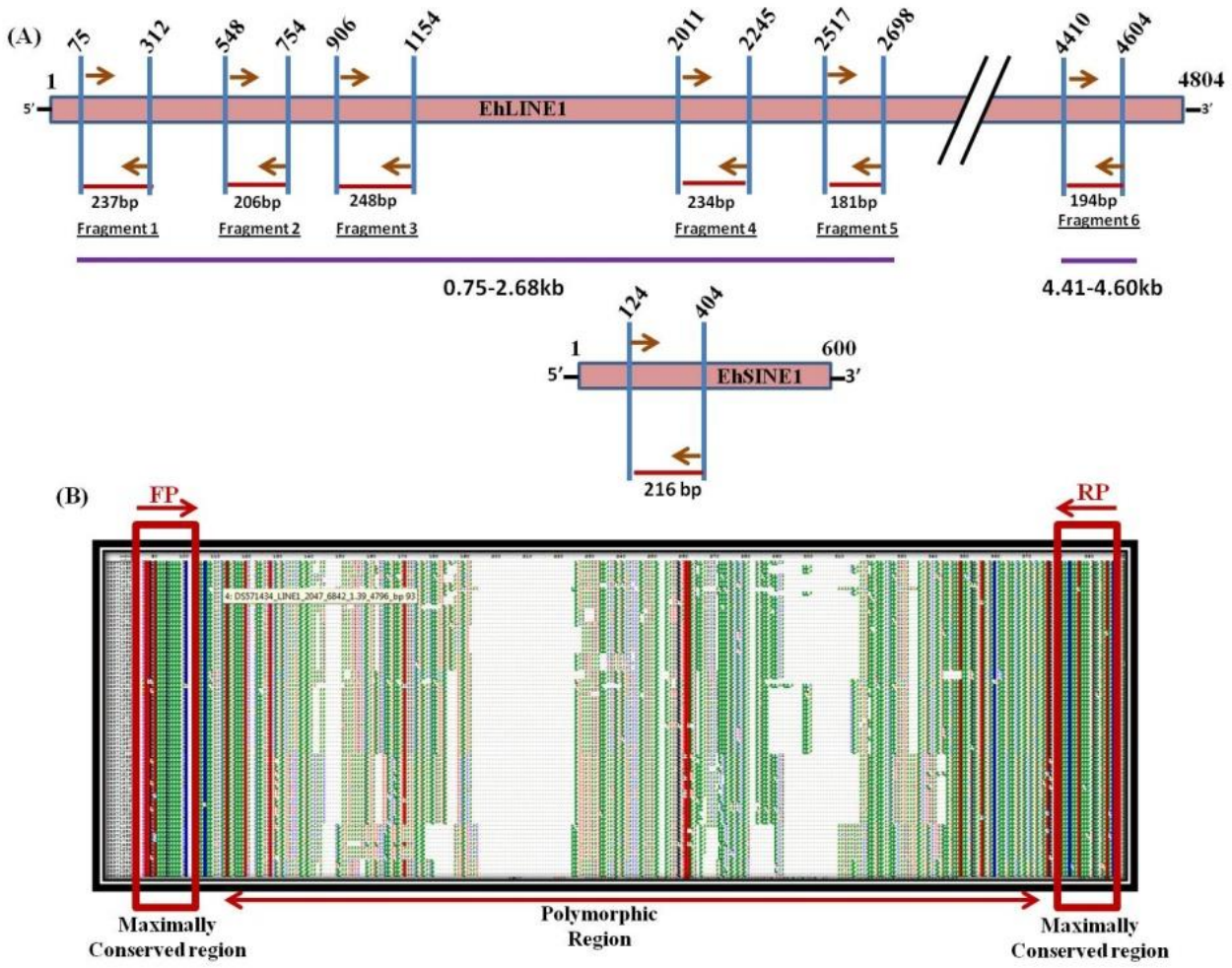


Figure 13: Schematic representation of Primer designing for targeted sequencing: (A) Six different regions in EhLINE1 to cover the full-length LINE1 and one region in EhSINE1 was selected for reverse transcriptase (RT)-PCR followed by targeted sequencing. (B) Primers were designed from the highly conserved regions in all the full-length EhLINE1 and EhSINE1 copies so as to amplify the maximum number of expressing copies.

Amplicons (DNA/RT-PCR)	Size (bp) (without adapter)	Size (bp) (with adapter)
LINE1Fragment 1	230	303
LINE1Fragment 2	206	272
LINE1Fragment 3	248	314
LINE1Fragment 4	234	300
LINE1Fragment 5	181	247
LINE1Fragment 6	194	260
SINE1	216	282

Table 3: Sizes of different amplicons of EhLINE1 and EhSINE1 with and without Barcode and adapter sequences.

For the expression analysis, cDNAs were prepared using total *E. histolytica* RNA and the reverse primers from each selected region (primer sequences are given in the appendix). Amplicons were obtained using the indicated primer pairs with cDNA as template. The amplified products were checked by agarose gel electrophoresis. Each band obtained by RT-PCR would be a mixture of sequences from all the amplified copies. As a positive control, amplicons were also obtained using DNA from a cloned copy of EhLINE1 which would have a unique sequence (DNA PCR lanes) (Fig.14).

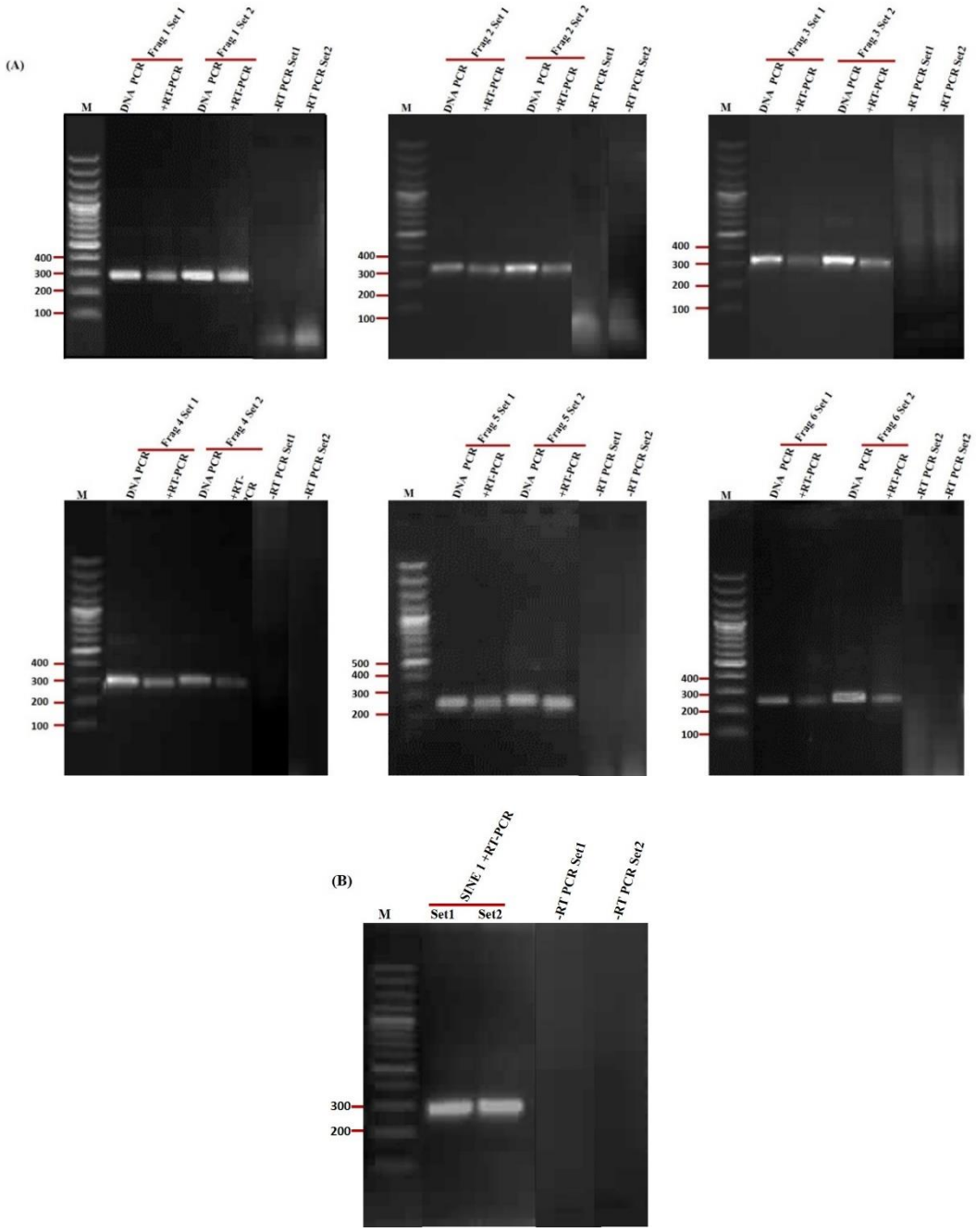


Figure 14: Agarose gel electrophoresis of LINE1 and SINE1 amplicons. (A) LINE1 amplicons were PCR amplified from EHLINE1 cloned DNA and cDNA template (two biological replicates) with set1 and set2 primers

having barcode and adapter sequences to check the specific amplification. The two sets of primers used had different adapter sequence in its forward and reverse primer so that it can be sequenced with two different adapter sequences from both the directions. -RT control and 100bp DNA ladder are shown. (B) EHSINE1 amplicon PCR amplified from cDNA template along with both the sets of primers and -RT control. All the amplicons checked on 1.5% agarose gel showed specific amplification of expected size (mentioned in table 3).

All the amplicons were sequenced using Ion Torrent platform. Two biological replicates were used for each sample (Set 1 and 2 in Fig.14). The sequences obtained were assigned to specific EhLINE1 copies based on 99% unique mapping. Analysis of the sequencing result showed that out of 967 EhLINE1 copies in the *E. histolytica* genome, 337 were expressed while the rest were silent. We found that all 57 full-length copies were expressed, as sequences corresponding to all of them were scored in the Ion Torrent output. Of the 910 truncated copies, 280 were expressed. In the DNA control sample, only the expected sequence of cloned EhLINE1 copy was obtained. In EhSINE1 expression analysis, we found that 84 copies out of the total 393 full-length copies were expressed. A copy was assigned as expressed on the basis of read count and considered as expressed if it had >10 read counts. The expressed copies had varying no of reads alignment ranging from 10 to 15000 read count. Of the 84 expressed copies, maximum reads were mapped with the SINE1 of R2 class followed by R4, R3, and R1 class, showing that R2 repeat containing SINE1 are highly expressed. The r4 class showed expression of all 3 copies. We could not get reads from the R1only, R3 only and None repeat containing SINE1classes.

Since the above analysis was done with PCR amplicons, one cannot comment on the relative expression status of the copies from number of reads as the PCR was not quantitative. This data was useful to differentiate the totally silent copies from transcriptionally active ones.

Further, we undertook a more precise transcriptome analysis by whole genome RNA-Seq with Poly(A)$^+$ RNA, in three independent biological replicates. RNA sequence data were obtained by paired-end deep sequencing using Illumina platform. Of the 35 million reads, >90% aligned with *E. histolytica* genome. In QC check, low-quality reads were trimmed by Trimmomatic. After trimming, sequences were mapped with EhLINE1 and EhSINE1 reference sequences and the expression was calculated by RSEM (RNA Sequencing using Expectation Maximization) considering a copy expressed if the read count was >10 (Detailed analysis of these data has been done by Devinder Kaur, Ph.D. scholar, salient features are summarized here). More than half of the reads aligning to the EhLINE1 sequence (72% on the basis of RT specific expression) came from full-length EhLINE1 copies. The contribution of truncated copies to the EhLINE1 transcriptome was less than that of full-length copies. The number of expressed EhLINE1 copies from RNA-Seq analysis was smaller than that obtained by the PCR approach (targeted sequencing). 57/967

EhLINE1 copies were expressed while the rest were silent (Table 4). Of the expressed copies 20 were full-length. In contrast, all 57 full-length EhLINE1 copies showed expression in targeted sequencing. The 37 full-length copies that were missed in RNA-Seq were scored in the very low expression category and were thus below the cut-off. Obviously, their expression level was sufficient to be detectable after RT-PCR. In RNA-Seq data, a total of 149 EhLINE1 copies scored <10 read count in both biological samples, and 108 copies had <10 read count in at least one biological sample. These showed expression in targeted sequencing, accounting for the discrepancy in the two data sets.

On the other hand, in EhSINE1 the number of expressed copies was higher in RNA-Seq data (157/393 full-length copies) compared with targeted sequencing (Table 4). We also found 6 out of 100 truncated EhSINE1 copies that showed expression. The copies that were missed out in ion torrent data had low Transcripts per million (TPM) values, between 2-5. It is possible that due to their low abundance and/or secondary structure, they were not efficiently reverse transcribed.

EhLINE1	**Total copies**	**Targeted amplicon sequencing**		**Illumina RNA sequencing**	
		Expressed copies	**Percentage (%)**	**Expressed copies**	**Percentage (%)**
Full Length	57	57	100	20	35
Truncated	910	280	30.7	37	04
EhSINE1 class	**Total copies**	**Expressed copies**	**Percentage**	**Expressed copies**	**Percentage**
R1	158	21	13.29	52	33
R2	67	55	82	61	91
R3	7	5	71.4	6	86
R4	3	3	100	3	100
R1 only	6	0	0	3	50
R3 only	89	0	0	12	13
None	63	0	0	14	22
Truncated	100	0	0	6	06

21.3

38.4

Table 4: Expression status of EhLINE1/EhSINE1 in Targeted sequencing /RNA-Seq data sets

Next, we checked the distribution of reads along the EhLINE1 sequence. The RT sequence showed very high read depth for almost all expressed copies, and these were located at approximately 2500 to 3800nt position. We found a total absence of reads from some regions of EhLINE1, especially a stretch between ORF1 and ORF2 in all copies (Fig.15). This is surprising, as LINEs are thought to

be transcribed into a continuous full-length transcript from an internal promoter at the 5'-end. A possible explanation for this observation is attempted in subsequent sections.

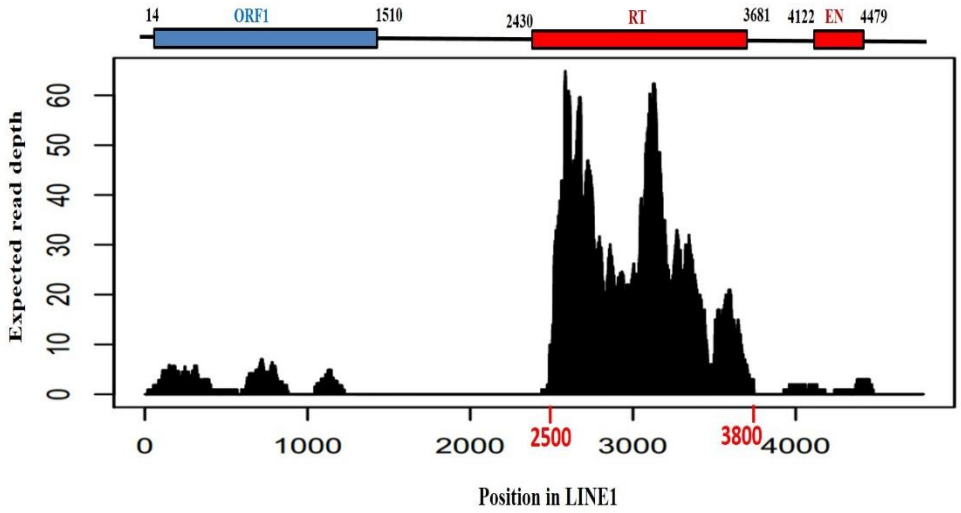


Figure 15: RNA-Seq reads alignment to a typical full-length EhLINE1 copy. Maximum reads mapped to the RT domain in ORF2. Much fewer reads mapped to theORF1 or the EN domain of ORF2. No reads mapped to the region between the two ORFs (analysis courtesy Kaur D, Ph. D. thesis).

3.4 Experimental validation of RNA-Seq data

To partially validate the expression pattern of individual EhLINE1 copies determined from RNA-Seq analysis, we obtained the expressed EhLINE1 sequences by RT-PCR of total RNA using primers from the conserved region of ORF1 5'-end (59-417) to obtain a 358bp amplicon. These primers would amplify a minimum of 92 EhLINE1 copies as judged from sequence identity. The amplicon was cloned into pGEM-T Easy vector and the inserts were sequenced from randomly picked clones. Sequence information was available for a total of 50 such colonies (table 5). The experimental data showed close correlation with the RNA-Seq data. The top 10 expressed copies in RNA-Seq includes 3 full-length and 7 truncated copies. All the 3 full-length copies were present in our sequenced colonies. The EhLINE1 sequence in scaffold DS571192 showed the maximum expression experimentally as it was present in 17/50 colonies. In RNA-Seq it ranked third with respect to expression level. The top most expressing copy in RNA-Seq (scaffold DS571495) was picked by 8/50 colonies and the second highest full-length copy in scaffold DS571290 was picked by 6/50 colonies. Two 5'-intact and 3'-truncated copies which were scored in RNA-Seq data were also picked up in our clone sequencing (table 5). Overall, our experimental result validated the data obtained from RNA-Seq.

Scaffold No.	Size	Identity with RT-PCR clones (%)	Number of clones	Log2 TPM (RNA seq. data)
DS571495	Full-length	99	8	6.4
DS571290	Full-length	100	6	5.1
DS571192	Full-length	100	17	3.7
DS571606	5' intact	99	2	2.6
DS571151	Full-length	99	3	2.5
DS571467	Full-length	98	4	2.1
DS571417	Full-length	99	2	2.09
DS571434	Full-length	98	2	1.8
DS571234	5' intact	99	1	0.5
DS571394	Full-length	100	1	-0.72
DS571505	Full-length	99	2	-1.7
DS571155	Full-length	100	1	-1.8
DS571156	Full-length	100	1	-2.2

Table 5: RT-PCR clone sequence analysis and its expression status in RNA-Seq data

3.4.1 Expression analysis of EhLINE1 by northern blotting

To check the physical status of EhLINE1 transcripts, northern blot analysis was done with total cellular RNA using probes from ORF1 and ORF2 (Fig.16A). Both probes gave a single distinct band of ~1.5kb. We also checked the transcription status in cells subjected to heat and oxygen stress. Again, both probes gave a single band of ~1.5kb and there was no significant change in the transcript levels in stressed cells. Importantly, no full-length transcript of 4.8kb was visible either in actively growing cells or in stressed cells (Fig.16B). To determine which regions of EhLINE1 correspond to the transcripts seen in northern blots, we used sub fragments of EhLINE1 as probes. The 3'-half of ORF1 coding region had already been used as a probe in a previous study (Yadav *et al*, 2012) and it hybridized with the same size band of ~1.5kb as shown here for 5'-end probe. Between the stop codon of ORF1 and the first AUG codon of ORF2 there is a 443bp 'spacer' region from which we designed probe S. Probes A to E covered the entire ORF2 coding region (Fig.16A). Northern data with these probes showed that the ~1.5kb transcript band came from the region covered by probes B, C and D. No signal was obtained from probes S, A and E (Fig.16C),

which was not due to poor labeling of the probes, as shown by dot blot hybridization with DNA of the corresponding probes.

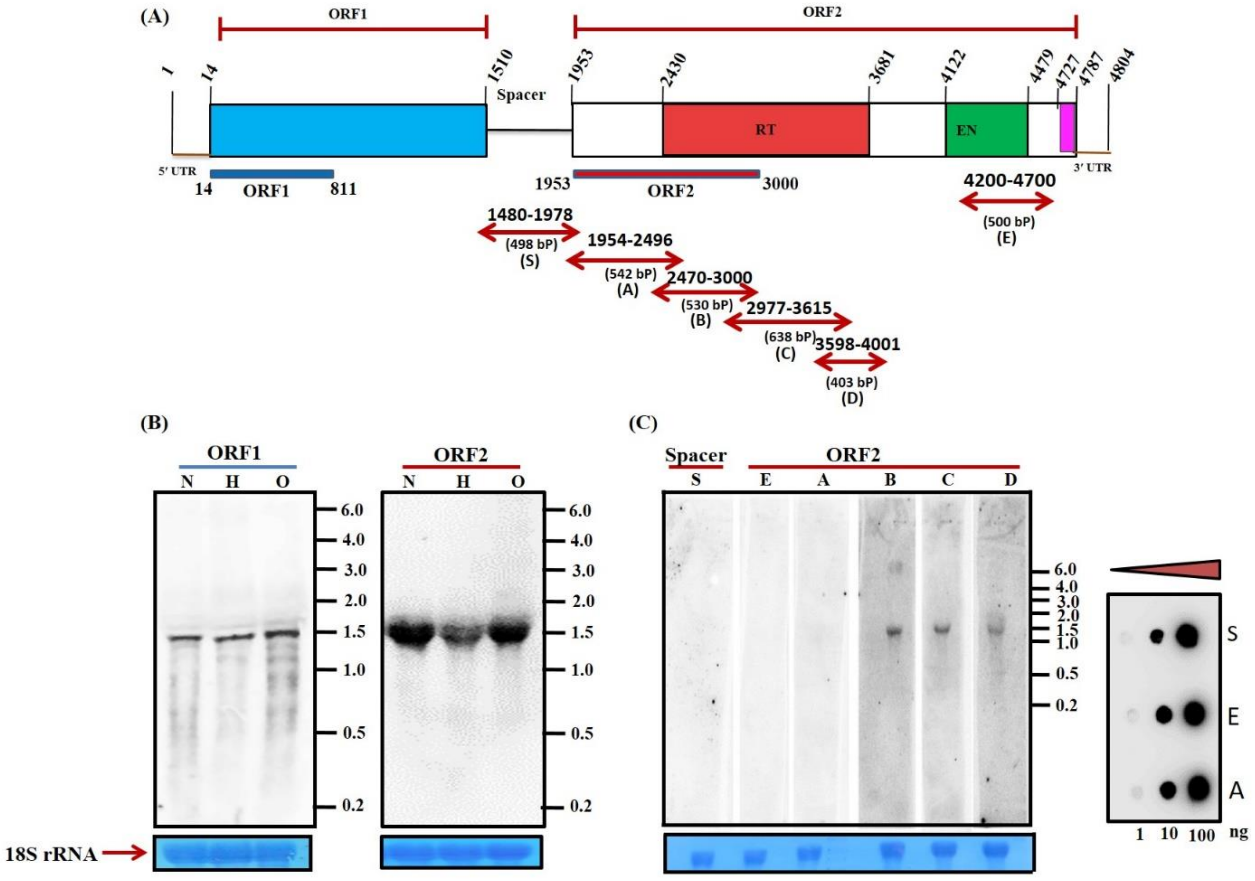


Figure 16: Northern blot with ORF1 and ORF2 probes. (A) Nt positions of probes from ORF1 and various regions of ORF2 are marked below. Probe 'S' covers the spacer region between the two ORFs. (B and C) Total RNA isolated from normal (N) and stressed cells [heat shock (H) and oxygen stress (O)] (details in method section) was electrophoresed, transferred to nylon membrane and hybridized with the indicated ORF1 and ORF2 probes. 18S rRNA was used as a loading control. Probes 'S', 'A' and 'E' did not give any signal in northern blot. Probe quality was confirmed by dot blot analysis with increasing concentration of DNA.

The northern data corresponded very well with the RNA-Seq data, as probes that failed to give any band in northern were from regions of EhLINE1 that gave negligible reads in RNA-Seq as well. The only exception was the EN domain region of ORF2 (covered by probe E) that gave no signal in northern hybridization but reads corresponding to this region were seen in RNA-Seq. Their level was comparable to reads from ORF1 region (Fig.15). It is possible that these reads come from very short transcripts of heterogeneous sizes and therefore are not seen as a sharp band in northern. The reads in our RNA-Seq analysis have an average size of 100 nucleotides.

Although 72% of the transcripts seen in RNA-Seq analysis of EhLINE1 came from full-length EhLINE1 copies, we failed to see any full length (4.8kb) transcripts in northern blots. It is generally believed that LINE elements are transcribed into polycistronic mRNAs from an internal promoter at the 5'-end (Macias *et al*., 2016; Heras *et al*., 2007)**.** Our results with EhLINE1 were thus surprising. There could be several explanations for this observation. (1) There may be multiple promoters and transcription termination sites in the EhLINE1 sequence giving rise to the shorter transcripts; (2) The full-length transcript of EhLINE1 may be rapidly processed co- or post-transcriptionally into shorter transcripts. (3) Short transcripts may arise from read-through transcription of truncated copies. The following experiments were done to seek an explanation for the northern data.

3.4.2 Does EhLINE1 contain a second internal promoter?

LINE elements are generally transcribed from an internal promoter located at the 5'-end. Earlier work from our lab had shown the presence of this internal promoter in EhLINE1 at the 5'-region. From deletion analysis, it was found that the promoter activity in a luciferase reporter assay was lost in fragments that included only 100bp from 5'-end (P-100) whereas a fragment that included 200bp showed promoter activity. Since we obtained only ~1.5kb transcripts in northern blots, we reasoned that transcription initiating at the 5'-end of ORF1 might terminate at its 3'-end, and there may be a second promoter downstream of ORF1 from where ORF2 transcripts might be initiated. To look for a promoter in this region we cloned a 1.5kb fragment (1511nt – 3000nt) spanning the 'spacer' between the two ORFs and including 1047bp (1953-3000nt) of ORF2 in p-Eh-Neo-LUC vector (replacing lectin promoter) upstream of luciferase (Fig.17A,B) (construct ORF2). *E. histolytica* trophozoites were transfected with this construct to check the expression of luciferase. A promoter less (P-less) construct, and parental vector with *E. histolytica* lectin promoter were used as negative and positive controls respectively. Activity of luciferase was measured in freshly prepared *E. histolytica* cell lysates using a kit from Promega. Results showed that luciferase expression with ORF2 construct was negligible and similar to P-less. Thus, we could not detect a second internal promoter in EhLINE1 downstream of ORF1 (Fig.17C). However, this needs to be confirmed by checking for the luciferase transcript as well.

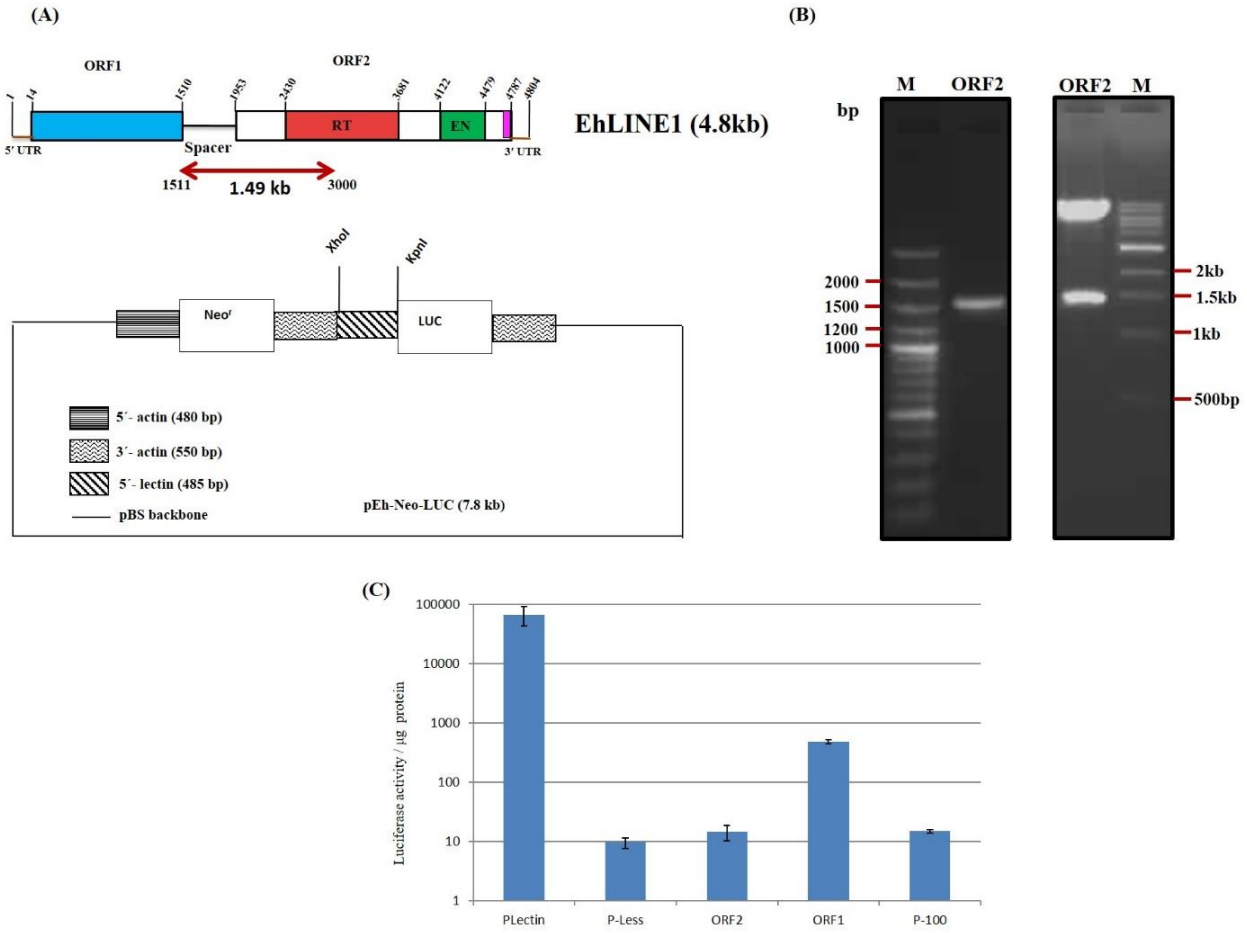


Figure 17: Looking for a second internal promoter in ORF2. (A) Primer position in EhLINE1 and map of the pEh-Neo-LUC vector. Primers were designed from the upstream of the RT domain to check the promoter activity in ORF2. (B) Mentioned 1.5kb region was PCR amplified; cloned into the pEh-Neo-LUC vector at the place of Lectin promoter. The clone was confirmed by restriction digestion showing expected 1.49kb band followed by transfection into normal proliferating trophozoites (C) Luciferase reporter assay was performed with the freshly prepared transfectant cell lysate as described in the method section. P-100 and P-Less used as negative controls in which promoter was deleted whereas ORF1 and P-Lectin (with parental lectin promoter) were used as positive controls. The data are average of three independent measurements.

3.4.3 The ORF2 transcripts originate from both full-length and truncated EhLINE1 copies from both directions

If a second internal promoter does not exist in EhLINE1, the 1.5kb transcripts corresponding to ORF2 could originate from the rapid processing of full-length EhLINE1 transcripts. Alternatively, they could arise from read-through transcription of truncated EhLINE1 copies lacking the ORF1 region. This was checked by analyzing the RNA-Seq data. It was found that ORF2 reads came from both full-length and truncated copies, with 72% reads coming from full-length EhLINE1 copies. Thus, the short size of ORF2 transcript cannot be explained by read-through transcription alone.

Moreover, the compact band of ~ 1.5kb seen in northern blots with ORF2 probe (in spite of heterogeneous nature of EhLINE1 genomic copies that are transcriptionally active), shows that the accumulated transcript is likely to arise from conserved processing events.

Further, we used strand-specific probes to check the orientation of transcripts seen in northern blots. Single-stranded DNA probes were generated by linear PCR of ORF1 and ORF2 templates using either reverse or forward primers along with α-P^{32}-labeled dCTP. Northern hybridization showed that in the case of ORF1, transcription was seen only in sense orientation whereas ORF2 transcripts were from both sense and antisense strands (Fig.18).

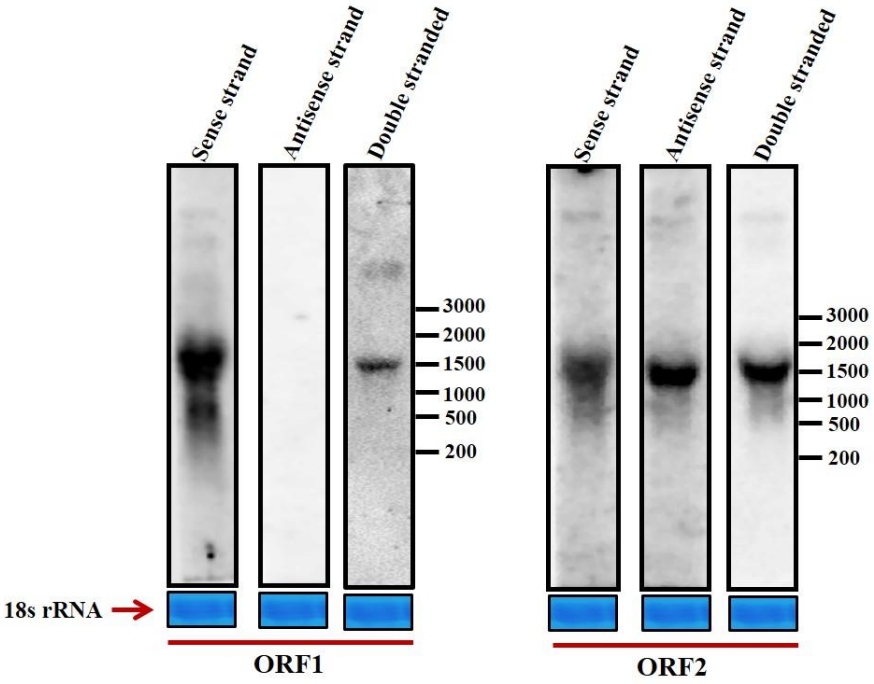


Figure 18: Northern blot analysis to check the direction of ORF1 and ORF2 transcripts. Sense and antisense transcripts were checked by using single stranded DNA probes for hybridization. 30µg of total RNA was electrophoresed, transferred to nylon membrane and hybridized. Single stranded DNA probes were generated by linear PCR with either reverse or forward primer. For ORF1 it was the 14-811 region while for ORF2 it was the B+C region (2470-3615). Double stranded probe was used for comparison and 18S rRNA as the loading control.

Both sense and antisense ORF2 transcripts were of comparable intensity, showing that both strands are actively transcribed. Further work needs to be done to check for the presence of a promoter at the 3'-end of EhLINE1 that may be responsible for antisense transcripts.

3.4.4 Polyadenylation status of EhLINE1 transcripts

Polyadenylation status of ORF1 and ORF2 transcripts was checked by RT-PCR of total RNA of *E. histolytica*, using oligodT primer for reverse transcription followed by PCR with ORF1 and ORF2-specific primer pairs as shown in Figure 19(A). To minimize non-specific reverse transcription with oligodT, a 45mer primer was used and RT reaction was performed at high temp (50^{o}C/1hr). The expected amplicon size for ORF1 was 249bp (position 905 to 1154) and ORF2 was 530bp (position 2470 to 3000). Both amplicons were obtained by RT-PCR (Fig.19B), showing that both ORF1 and ORF2 transcripts are likely to be polyadenylated. To further confirm the polyadenylation, we repeated the same experiment with poly(A)$^{+}$ RNA and obtained the same results (Fig.19C). U3 snoRNA was used as a negative control for RT with oligodT primer, as it is not polyadenylated. It did not give any amplicon when RT reaction was done with oligodT primer but gave the expected amplicon of 212bp with U3-specific reverse primer (Fig.19D).

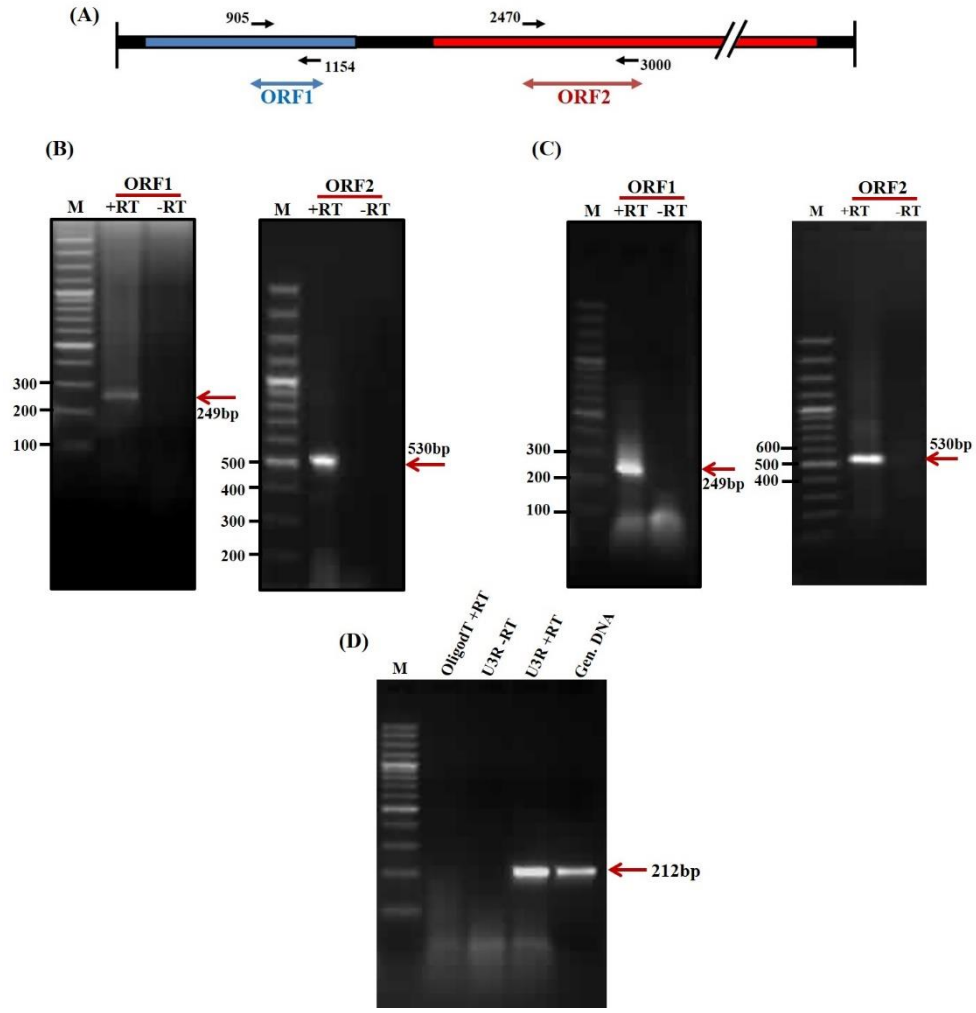


Figure 19: Polyadenylation of ORF1 and ORF2 transcripts. cDNA was synthesized using a long oligodT primer (45mer) at 50^{o}C, to minimize nonspecific binding as *E. histolytica* has highly AT rich genome. (A) Schematic

representation of primer pairs used for PCR of cDNA synthesized by oligodT. (B) DNase treated total RNA (5µg) was used for cDNA synthesis with oligodT primer, followed by PCR with ORF1 and ORF2 specific primers. Amplicons were checked by 1.5% agarose gel electrophoresis. (C) As in (B), but poly(A)$^+$ RNA (500ng) was used for cDNA synthesis. (D) To check the specificity of oligodT primer, the cDNA obtained from total RNA in (B) was used to amplify U3 snoRNA. cDNA was also synthesized with U3 specific reverse primer (U3R) followed by PCR with U3 specific primers. An amplicon of the expected size (212bp) was obtained only from cDNA made with U3R and not with oligodT primer.

The ORF1 transcripts seen in northern blots span the entire ORF1 coding region. In addition, the ORF1 polypeptide is also present in *E. histolytica* cells (Yadav *et al*. 2012). Hence most of these transcripts are likely to be capped and polyadenylated. However, the ORF2 transcripts correspond only to the RT domain and do not span the entire ORF2 region. The ORF2 polypeptide was also undetectable in *E. histolytica* using an antibody against the endonuclease domain. At present, attempts to raise antibodies against RT domain have been unsuccessful; hence the translation status of the ~1.5kb ORF2 transcript is unknown.

3.4.5 Locating the 3'-ends of ORF1 and ORF2 transcripts

To locate the 3'-end of ORF1 and ORF2 transcripts we mapped the location of poly(A) tail by RT-PCR using oligodT primer for reverse transcription, followed by PCR with oligodT as reverse primer and selected oligonucleotide sequences close to the 3'-end as forward primers. With ORF1 the forward primer located at position 1305 gave an amplicon of 267bp while the primer at position 1201 gave an amplicon of 372bp, as estimated from the migration of bands in 100bp ladder (Fig.20A). The difference in size of the two amplicons (105bp) corresponded well with the distance between the positions of the two forward primer FP1 and FP2 (104bp) (Fig.20B). From this (deducting 45nt of the oligodT primer) the approximate location of 3'-end of ORF1 would be at position 1528 (wrt +1 of EhLINE1). Similarly, with ORF2 the forward primer located at position 3814 gave an amplicon of 211bp while the primer at position 3598 gave an amplicon of 429bp (Fig.20A). The difference in size of the two amplicons (218bp) was in agreement with the distance between the two forward primers (216bp) (Fig.20B). From this, the approximate location of 3'-end of ORF2 would be at position 3982 (wrt +1 of EhLINE1). This shows approximate correlation with the end of reads from RT region which is at 3800nt position as seen by RNA-Seq (Fig.15).

Assuming that ORF1 transcription would initiate from the start of EhLINE1, and it's 3'-end is located at 1528, one would predict the transcript size to be 1528nt plus the length of poly(A) tail which is estimated to be about 25nt (Hon *et al*., 2013). The observed size of this transcript from northern blots is ~1.5kb, which is in agreement with the predicted size.

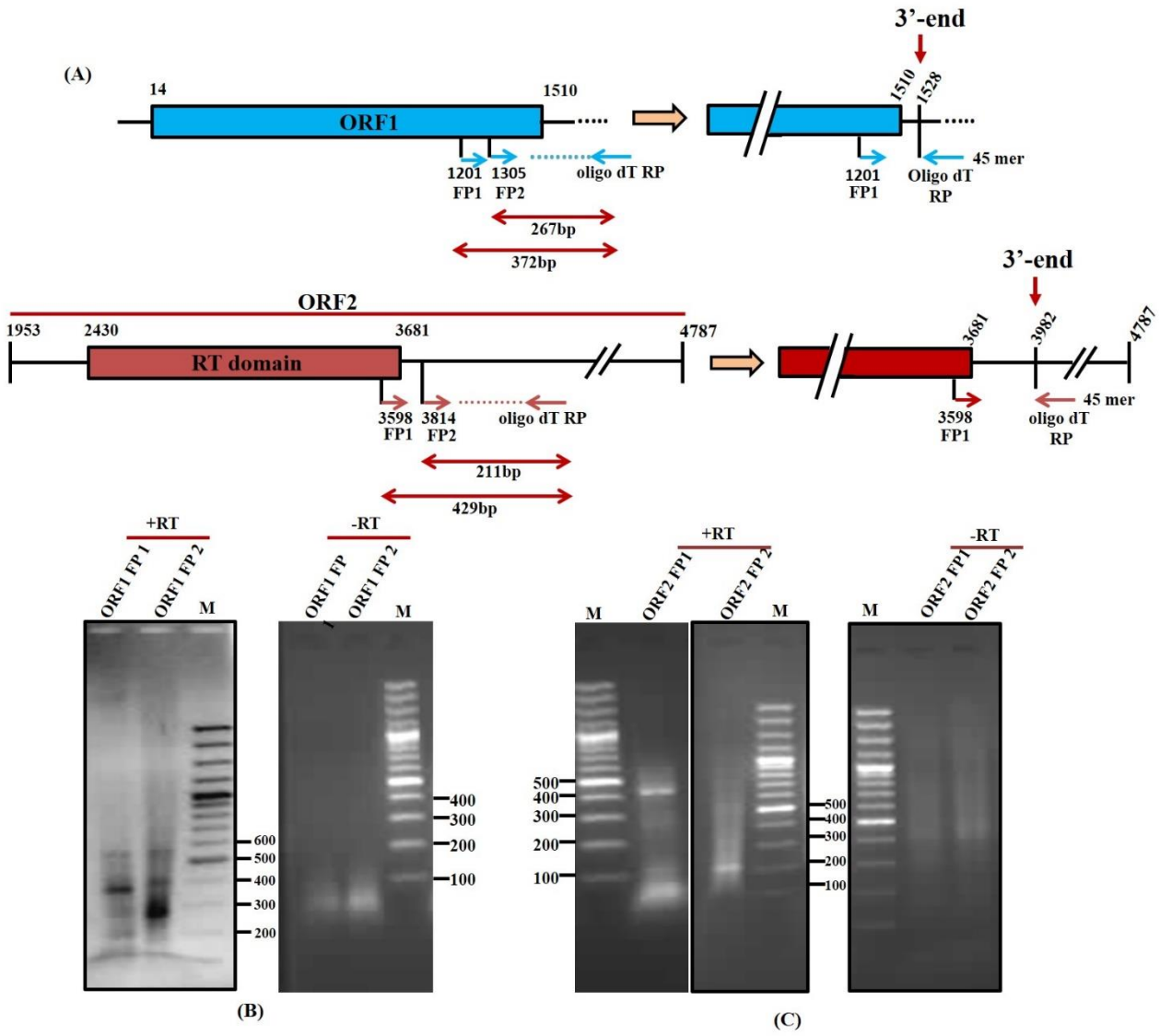


Figure 20: Determination of 3'-end of ORF1 and ORF2. (A) Schematic representation of forward primers used in PCR. To check the 3'-end of ORF1, two forward primers spaced 104bp away from each other were used. For ORF2 the two forward primers were spaced 216bp away. (B and C) cDNA was synthesized with oligodT followed by PCR with the indicated forward primers along with oligodT as reverse primer. Products were checked on 1.5% agarose gel. No specific amplification was seen in –RT PCR control.

3.4.6 5'-end mapping of ORF2

In the previous sections, we have shown that the predominant ORF2 transcript is ~1.5kb in size and its location from RNA-Seq data is between nt 2500 to 3800 (Fig.15). To experimentally validate the 5'-end predicted from RNA-Seq, we did primer extension using a primer with 3'-end located at nt position 2622 (primer sequence is given in appendix) (Fig.21A). Primer extension was performed with total RNA and end labelled primer as mentioned in Methods section. A product of 100nt was visible in +RT lane whereas no product was seen in -RT (negative control) (Fig.21B). The

approximate 5'-end of this transcript was thus mapped to nt position 2522, which corresponded well with the RNA-Seq data in which the reads from RT region started from nt position 2500.

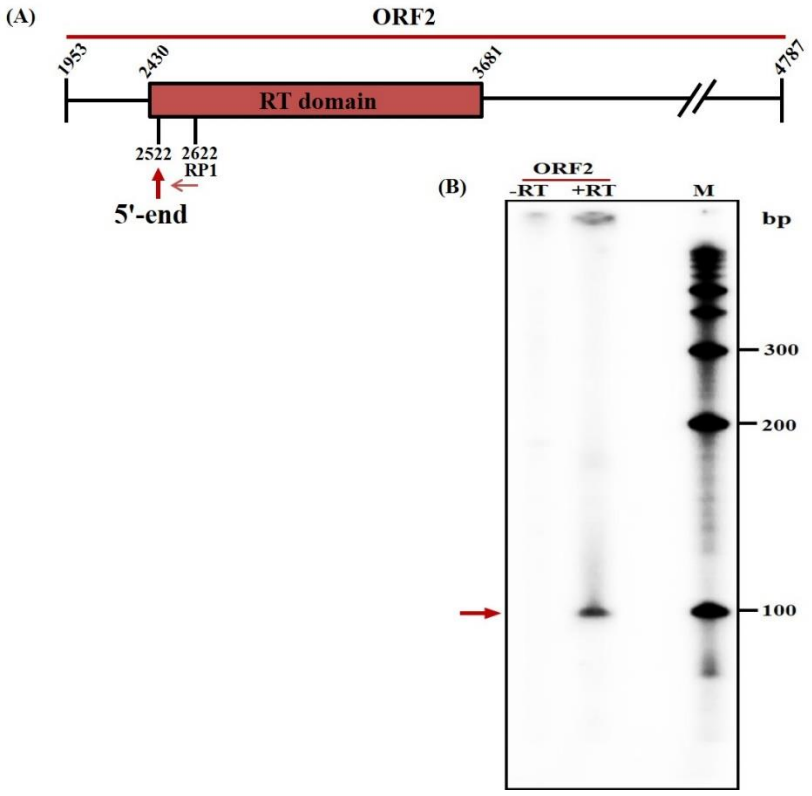


Figure 21: Primer extension to map 5'-end of the ORF2 transcript. End labeled reverse primer (3'-end at nt position 2622) was used to perform primer extension along with total RNA and revertaid RT (Fermentas) as per manufacturer protocol. The reaction was incubated at 42ºC/1hr followed by heat inactivation at 70ºC/10min. The product was checked on 12% denaturing UREA-PAGE with labeled 100bp Ladder. -RT used as a negative control.

From the above results the approximate 5' and 3'-ends of ORF2 transcript have been mapped experimentally at nt position 2522 and 3982 respectively. (The same by RNA-Seq is 2500 and 3800 respectively). The transcript size from experimental data is 1460nt plus the length of poly(A) tail (~ 25nt). This correlates well with the observed size of ~1.5kb in northern blots.

On the basis of these results, we can conclude that two independent transcripts of 1.5kb corresponding to EhLINE1 ORF1 and ORF2 exist in *E. histolytica*. Since they arise from full-length copies, they might be derived from post transcriptional processing. Further studies are needed to characterize the antisense transcript from ORF2; and to detect any read through transcription of EhLINE1 sequences, especially truncated copies, from neighboring genes.

3.5 Methylation status at promoter region of EhLINE1

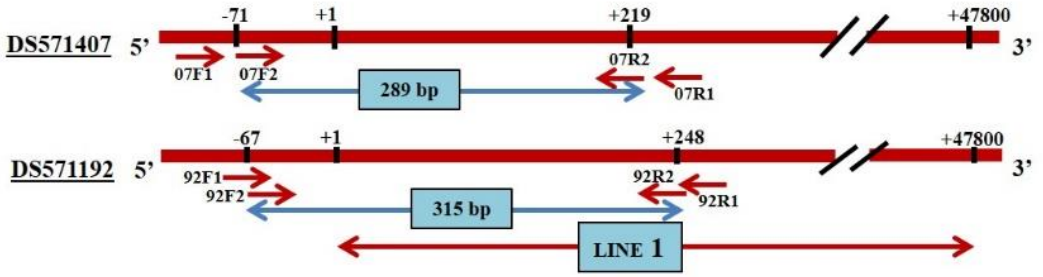
In mammals, a principle mechanism for retrotransposon silencing in both germ cells and somatic cells is transcriptional repression through DNA methylation. Mostly methylation takes place at the cytosine residues in the context of CpG dinucleotides. Inhibition of DNA methyltransferases leads to increased expression of retrotransposons and endogenous retroviruses (Ollinger, 2010). Transposon promoters are inactive when methylated, and suppression of their expression appears to be a primary function of cytosine methylation (Yoder, 1997; Hackett, 2012). Since promoter DNA methylation has been correlated with transcriptional silencing in model systems (Crichton, 2014) we were interested to know whether the 5'-end of EhLINE1 (where the internal promoter is located in LINE elements) showed cytosine methylation and whether this correlated with transcription status of individual EhLINE1 copies.

3.5.1 Cytosine methylation status of the promoter region of transcriptionally active and silent EhLINE1 copy

From the validation of expression data (shown in section 3.4) the EhLINE1 copy in scaffold DS571192 was selected for further analysis as an expressed copy, since it ranked amongst the top10 EhLINE1-expressed sequences in RNA-Seq; it was present in the 50 colonies sequenced from RT-PCR; and it showed the closest match (96% identity) with the consensus EhLINE1 sequence (Bakre *et al*., 2005). For silent copy, we selected the EhLINE1 sequence in scaffold DS571407 since it showed zero expression in RNA-Seq data, was not scored in the RT-PCR analysis (its sequence matched completely with the RT-PCR primers used), and was full-length. Methylation status of the 5'-end of these two copies was checked by bisulfite (BS) treatment of total genomic DNA, which converts unmethylated cytosines to uracil in DNA, while methylated cytosine remains protected (Fraga *et al*., 2002). The individual copies were amplified using locus specific upstream primers (92/07F1 and R1). Nested primers were used to get specific amplicons (92/07F2 and R2). Both amplicons had a total of 36 cytosines. The 289bp amplicon from the silent copy contained 218bp of EhLINE1 sequence with 2 CpG sites, while the 315bp amplicon from the expressed copy contained 248bp of EhLINE1 sequence, with 3 CpG sites (Fig. 22A). BS-converted primers (C–T) only amplified BS-treated DNA, and vice versa, showing that the bisulfite treatment was successful (Fig. 22B).

(A)

S No.	Accession No.	LINE1 Position	Length	Expression status	Expected amplicon size	No. of Cs /CpG in amplicon
1	DS571407	14251-19035	4.78kb	Silent	289bp	36/2
2	DS571192	10681-15465	4.78kb	Expressed	315bp	36/3



(B)

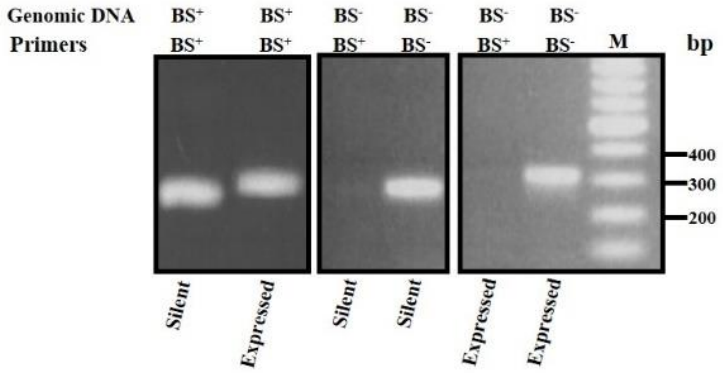


Fig. 22. Cytosine methylation status of the promoter of transcriptionally active and silent EhLINE1 copies. (A) The expressed and silent EhLINE1 copies selected to check cytosine methylation status at their promoter site are shown. Bisulfite converted (BS^+) and normal (BS^-) primers, including nested primers (07F1, F2, R1, R2 series for silent copy and similar 92 series for expressed copy), for bisulfite PCR (BS-PCR) were designed from the locations shown. The forward primers were upstream of EhLINE1 so as to amplify the specific EhLINE1 copy. (B) BS-PCR of expressed and silent EhLINE1 copies with BS-treated and untreated genomic DNA to confirm the primer specificity. The BS^+ primers amplified only the BS-treated DNA and vice versa.

Amplicons were sequenced to determine the extent of cytosine methylation. Sequence analysis showed that all the cytosine residues in DNA were converted to thymine upon treatment with bisulfite, showing that none of these cytosine residues were methylated in either of the two copies. Sequence alignment of the 5'-region of selected silent and expressed EhLINE1 copies shows that the two copies share 89% sequence identity across their entire length (Fig. 23).

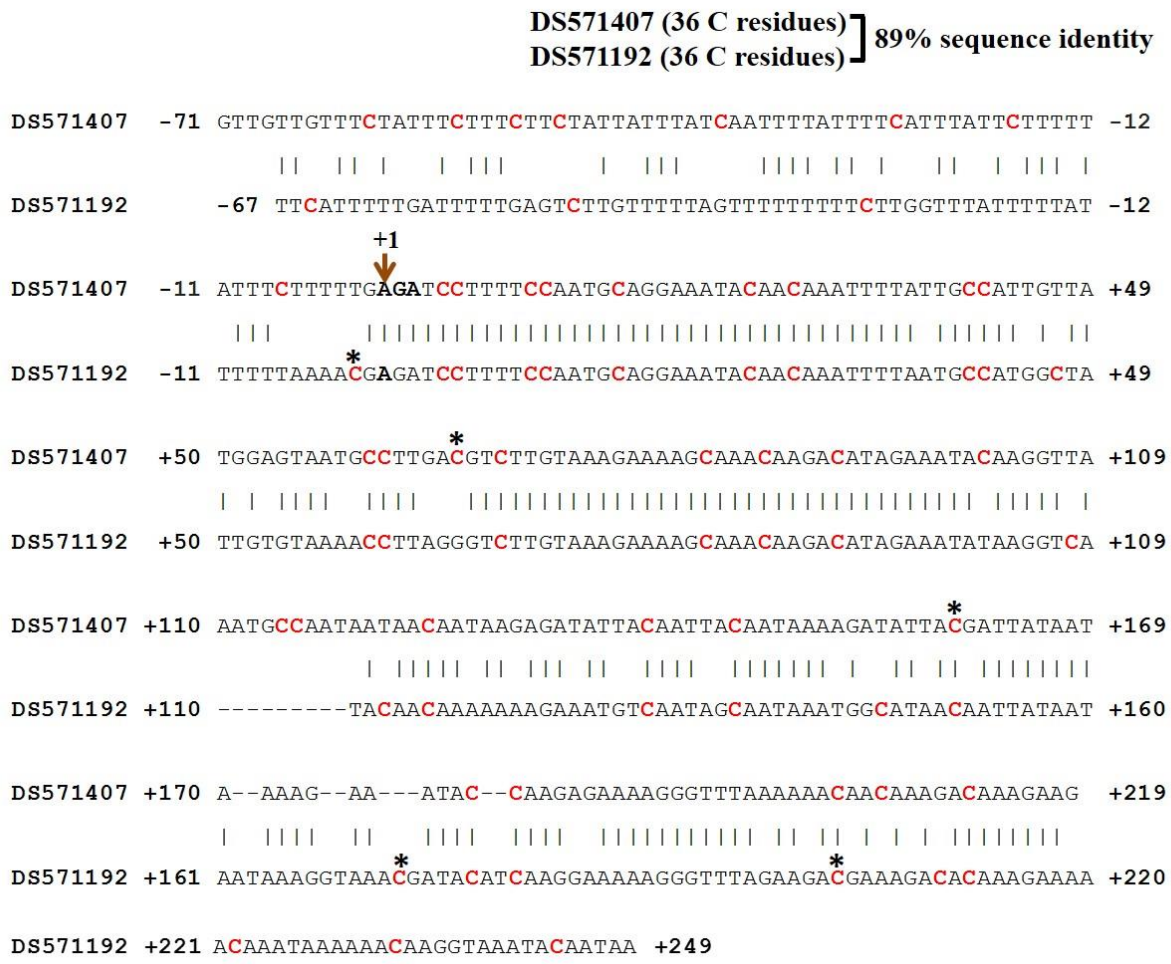


Fig 23: Nucleotide sequence alignment of the 5'-region of silent and expressed EhLINE1 copies selected for analysis. +1 indicates the start of EhLINE1 sequence. The two copies share 89% sequence identity across their entire length. All the C residues which were converted to T after bisulfite treatment are shown in red. The CpG residues are marked by an asterisk.

To show that the absence of cytosine methylation was not due to a technical problem we introduced methyl residues at CpG sites by treating genomic DNA with CpG Methyltransferase (M.SssI), followed by bisulfite conversion and amplicon generation as described above. The expressed EhLINE1 copy in scaffold DS571192 has 3 CpG sites. Their methylation status was checked in M.SssI-treated DNA, and 2 out of the 3 cytosines were protected from bisulfite conversion, showing that the observed absence of cytosine methylation in genomic EhLINE1 copies was not likely to be an experimental artefact. (All 3 cytosines may not be protected due to incomplete methylation by M.SssI).

3.5.2 Detection of cytosine methylation at selected sites in a larger subset of EhLINE1 copies

In the above experiment, we looked at methylation of all cytosines in the 5'-regions of only two EhLINE1 copies. We next determined the methylation of a few selected cytosines but in a larger subset of EhLINE1 copies. For this, we adopted single nucleotide incorporation assay approach using end labeled primer (details are given in Materials and Methods section). The assay was standardized by looking at incorporation of selected nucleotides by providing only the next complementary nucleotides to the primer-template in separate reactions (Fig.24A). Optimum dNTP concentration was determined for the +7nt reaction and found to be 40µM (Fig.24B). The +1, +2 and +7nt incorporations were checked by providing respective complementary nucleotides and the expected size band was obtained (Fig.24C).

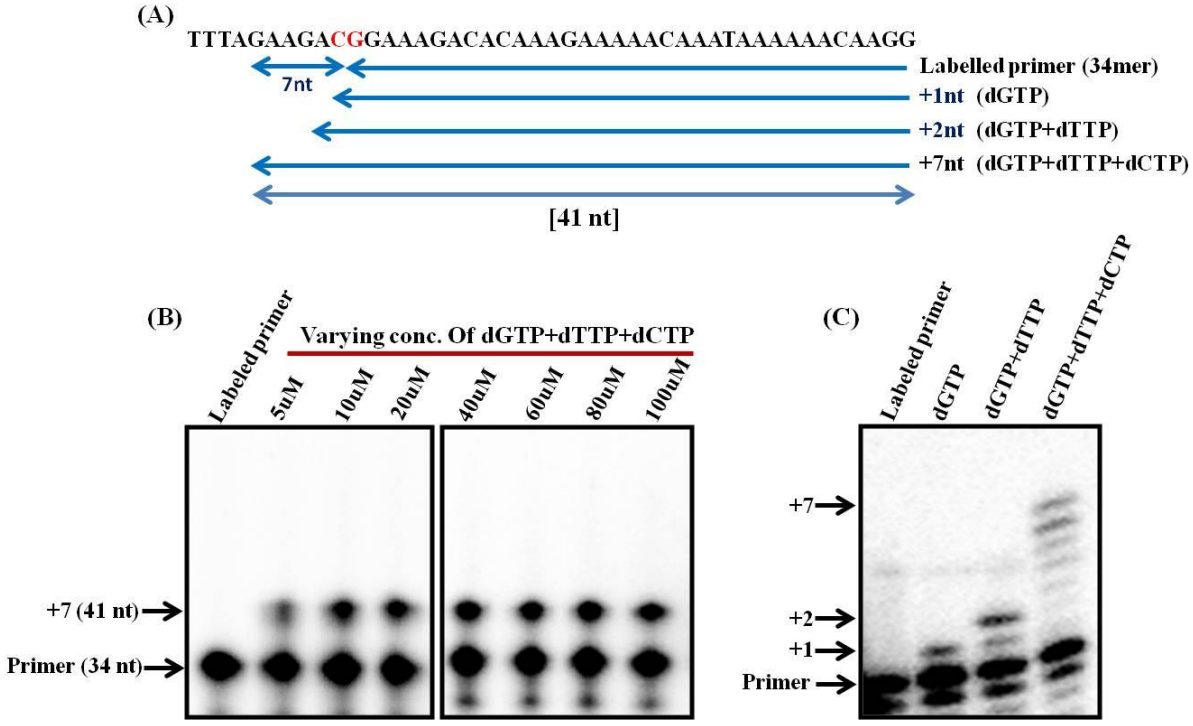


Figure 24: Standardization of nucleotide incorporation. (A) Labeled primer (34nt) was used along with appropriate dNTPs to obtain +1, +2 and +7 nucleotides incorporation. (B) Optimum dNTP concentration was determined by doing the reaction at different concentrations and products were checked on 7M 10% Urea-PAGE. 40µM concentration showed optimum incorporation. (C) Incorporation increased the growing complementary strand by +1, +2 and +7 nucleotides. Nucleotide shift was checked on 7M 6% sequencing gel. Few bands below the labeled primer were truncated primer as it was not PAGE purified.

To determine the methylation of selected cytosines in a larger subset, a few cytosines were selected by aligning all 5'-intact copies (256, including 57 full-length copies) of EhLINE1and looked for conserved CpG sites in the first 500bp. Four such sites were found to occur frequently, with at least one of the four sites present in 160 copies. Primers were designed from conserved sequences flanking these sites to obtain two amplicons containing two CpG sites each. The amplicons were obtained from bisulfite-treated and untreated genomic DNA using bisulfite-converted and non-converted primers respectively as described above. For each CpG site, a complementary primer was used with its 3' ending at C (complementary to the G residue in CpG). Primers were also designed to score three non-CpG sites in amplicon 1. These were present in a larger number of copies, with at least one site present in 200 copies (Fig.25A). The end-labeled primers were annealed with the amplicons from bisulfite-treated DNA and allowed to incorporate a single nucleotide (A or G), which would reflect the methylation status of the cytosine at that site. Amplicons from non-bisulfite treated DNA obtained with non-bisulfite converted primers were used as a control. The data showed that dATP was incorporated with bisulfite-treated DNA, while dGTP was incorporated with untreated DNA at all seven sites, showing lack of extensive cytosine methylation at these sites (Fig. 25B). If a small subset of the copies were methylated, their number could be estimated by determining the ratio of radioactivity at the +1-position compared with origin in the dGTP lane, by densitometry. This ratio was close to zero for dGTP in all samples, showing negligible levels of methylation.

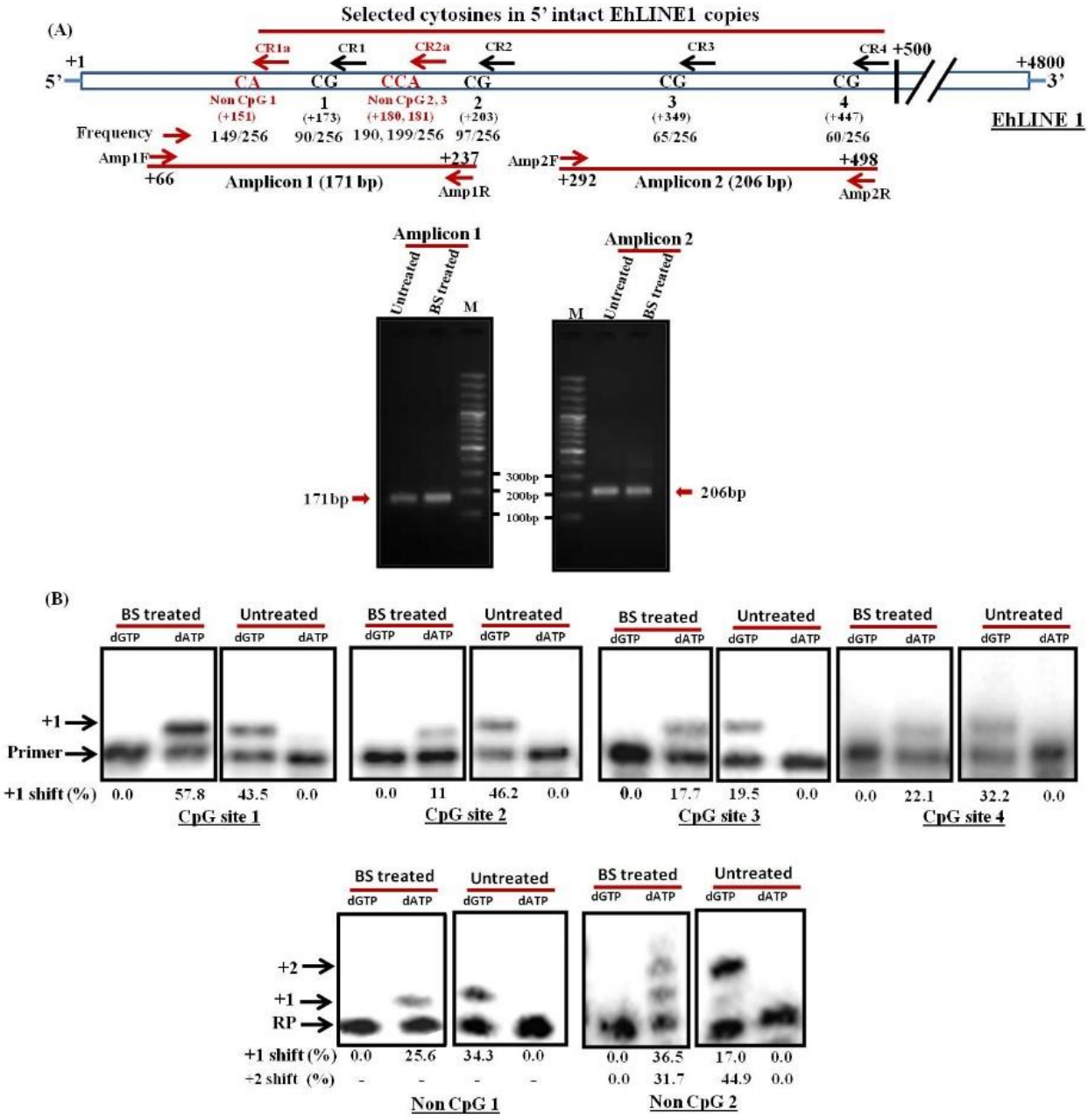


Fig. 25. Detection of cytosine methylation at selected sites in the promoter of a large subset of EhLINE1 copies. (A) The location of four conserved CpG sites and three Non-CpG sites in the 5–500bp region of EhLINE1 copies are shown, along with positions of the two amplicons containing these sites. Of the 256 5'-intact (including 57 full-length copies) EhLINE1 copies the number of copies in which the selected CpG residues occur is indicated below each CpG site. Reverse primers CR1, 2, 3 & 4 (shown above each site) were designed such that they end at the 'G' residue, to be used for single nucleotide incorporation assay opposite the 'C'. Non-CpG Reverse primers CR1a and CR2a were also designed with the same strategy. (B) Single nucleotide incorporation assay. Amplicons 1 and 2 were obtained from BS-treated and untreated DNA using primer pairs Amp 1F/1R and Amp 2F/2R respectively. Amplicon DNAs were annealed with respective end-labeled reverse primer for each site and extended in presence of either dGTP or dATP. +1 and +2 are the shift after nucleotide incorporation and percent +1 and +2 shift measured by densitometry is indicated.

3.5.3 The promoter of *E. histolytica* HSP70 gene remains methylated during heat shock when transcription is up regulated

Our data with EhLINE1 copies showed negligible cytosine DNA methylation at the sites examined by us in both expressed and silent copies, indicating that DNA methylation is unlikely to be involved in transcriptional regulation of these elements. To see whether the lack of correlation of cytosine methylation with transcription status was unique to EhLINEs, or was a more general feature of *E. histolytica*, we checked methylation of HSP70 whose promoter is fully methylated in normal *E. histolytica* cells (Fisher *et al*., 2006). The methylation status of this gene has not been checked during heat stress, which could directly correlate transcriptional control of this gene with promoter methylation. We used the methods described above for EhLINE1 to check the methylation of HSP70 promoter region (−201 to +42) both under normal and heat-stressed (42^{o}C for 60min) conditions. Sequencing of bisulfite-treated DNA showed that all cytosines were methylated in both conditions (Fig. 26).

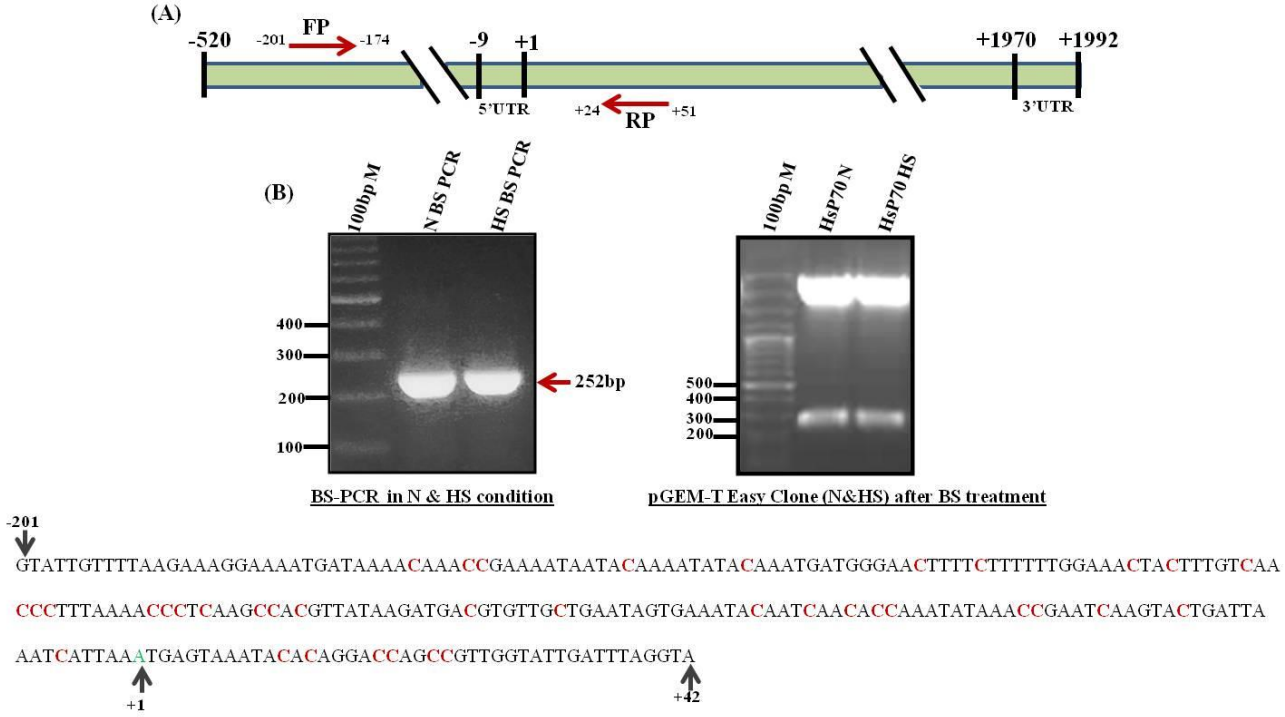


Fig. 26. Bisulfite sequencing of the promoter region of Hsp70 (EAL45068) gene copy shown to be methylated (Fisher *et al*., 2006) under normal (N) growth conditions. The same was also checked after heat stress (HS). (A) Positions of primers used to obtain the amplicons for BS-sequencing are shown. (B) BS-PCR amplicons were cloned into pGEM-T Easy vector followed by Sanger sequencing. The sequencing result showed methylation of all the 'C' residues in both N and HS conditions.

We also checked methylation of two selected CpG sites by incorporation of dGTP/dATP, which again showed that both sites were methylated in normal and heat-stressed cells**.** As a control, the EhLINE1 showed no methylation in the same DNA samples (Fig. 27).

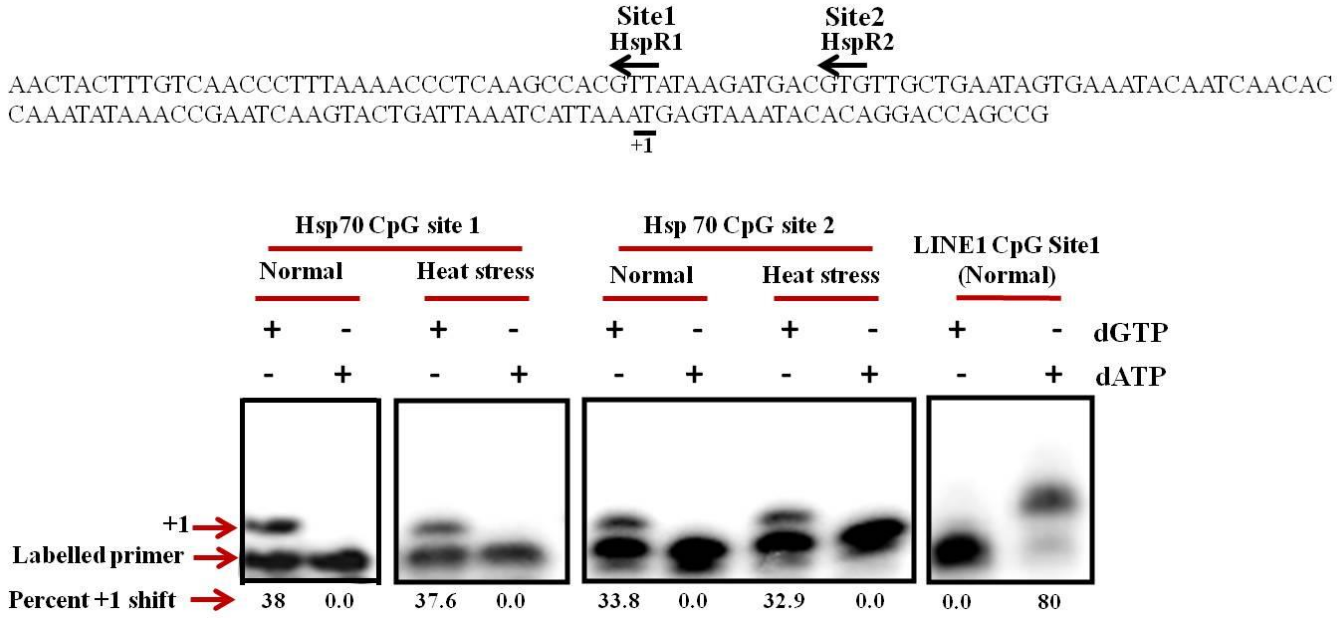


Fig. 27. Single nucleotide incorporation assay in the Hsp70 gene copy (EAL45068) at selected CpG sites: Two CpG sites were selected in the region known to be methylated (Fisher *et al.*, 2006) and primers ending at 'G' were made. DNA obtained from normal and heat-stressed cells (42ºC/60min) was treated with bisulfite, annealed with primers, and the assay was done as described for EhLINE1 in Fig. 25. In the same assay, DNA from non heat-stressed cells was used for EhLINE1 CpG site 1, which showed opposite results compared with Hsp70.

Further, we looked at expression status of HSP70 in heat-shocked cells by northern hybridization. Transcript levels were negligible in normal cells and as expected transcription increased to high levels upon heat shock (Fig. 28A). *E. histolytica* has 17 copies of HSP70 gene reported in the data base. To specifically determine transcript levels of the gene copy that has been used for cytosine methylation analysis, we used gene-specific primers for quantitative RT-PCR, which showed 8.5-fold upregulation of this copy upon heat shock (Fig. 28B). Since the cytosine methylation status of the HSP70 gene promoter remained unchanged although its transcription increased tremendously, the data directly demonstrate that DNA methylation was not involved in transcriptional regulation of this gene also, and this lack of correlation is likely to be a more general feature of *E. histolytica* genes.

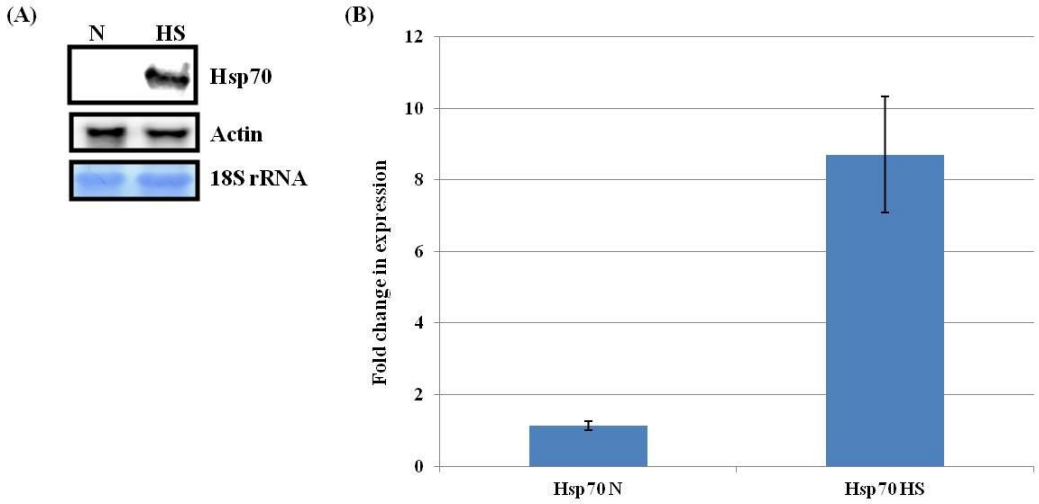


Figure 28. Expression of Hsp70 gene in normal (N) and heat-stressed (HS) conditions. (A) Northern blot analysis with Hsp70 probe in N and HS conditions. Actin was used as a control. 18S rRNA used as the loading control. (B) The expression level of the Hsp70 copy (EAL45068) used for DNA methylation analysis, quantified by qRT-PCR with primers qHspF and qHspR showed 8.5 fold increase in transcript levels in HS cells.

On the basis of above results, we can state that EhLINE1 promoter sequences are almost devoid of cytosine DNA methylation, and there may be little correlation between cytosine DNA methylation at promoter regions and transcription status in *E. histolytica.*

3.6 Overexpression and purification of EhLINE1 ORF2p

3.6.1 Expression and purification of ORF2p in bacterial system

Previously in our lab, the full-length EhLINE1 was reconstituted which contained complete reading frames of both ORF1 and ORF2. ORF1 and ORF2 were also cloned separately and tried to overexpress in various *E. coli* expression vectors. ORF2p is known to express at a significantly lower level (Dai *et al*, 2014), and has been very difficult to purify from *E. coli.* Along with its overall low expression, it also showed degradation during expression. Codon biasness is a problem for most amoebic proteins; so, we used codon plus *E. coli* host cells like BL21 (RIL) or *E. coli* (rosetta) for expression. We tried various host and vector combinations along with different expression conditions (summarized in Table 6) and were able to express full-length ORF2p (940aa) although at a low level in expression vectors pET30b, pET21a and pGEX-4T1 in the *E. coli* BL21 (RIL) strain. Full-length ORF2 was cloned in pET30b (*Kpn*I-*BamH*I site) and pET21a (*BamH*I-NotI site) with N-terminal and C-terminal polyhistidine (6xHis) tag respectively, followed by expression in BL21 (RIL) strain. Expression levels were similar in both the vectors (Fig.29A). Imidazole concentrations of 50, 80 and 100mM were used for the washing, and up to 100mM

imidazole the protein remained bound to the resin. In fact, bound protein showed poor elution from Ni-NTA agarose resin, and at 250mM imidazole concentration also we were not able to elute it efficiently from the resin. We analyzed the ORF2p sequence and found that it is highly cysteine rich; containing 14 cysteine residues, which could be the reason for poor elution. Use of 20mM DTT along with imidazole helped in getting the protein eluted (Kiedzierska *et al.*, 2008). However, both the input and the eluted material showed a lot of protein bands as revealed by western blot with the anti-his antibody (Fig29 B, C).

Expression vector	Associated Tag	Induction temp.	Induction duration	Inducer (IPTG)	Host (E coli.) strain	Expression status
pET30b	N-terminal polyhistidine (6xHis)	16°C 18°C	10 hour/O/N 6 hour 4 hour	0.5mM	Rosetta	No expression
					BL21 (RIL)	Low expression
					Shuffle	No expression
					Rosetta co-transformed Shuffle	No expression
					RIL co-transformed Shuffle	No expression
					P-rare co-transformed Shuffle	No expression
pET21a	C-terminal polyhistidine (6xHis)	16°C 18°C	10 hour/O/N 6 hour 4 hour	0.5mM	Rosetta	No expression
					BL21 (RIL)	Low expression
					Shuffle	No expression
					Rosetta co-transformed Shuffle	No expression
					RIL co-transformed Shuffle	No expression
					P-rare co-transformed Shuffle	No expression
pGEX-4T1	N-terminal GST	16°C 18°C	10 hour/O/N 6 hour 4 hour	0.2mM/0.5mM 0.5mM/1mM	Rosetta	No expression
					BL21 (RIL)	Low expression

Table 6: Various host and vector combinations and expression conditions tested

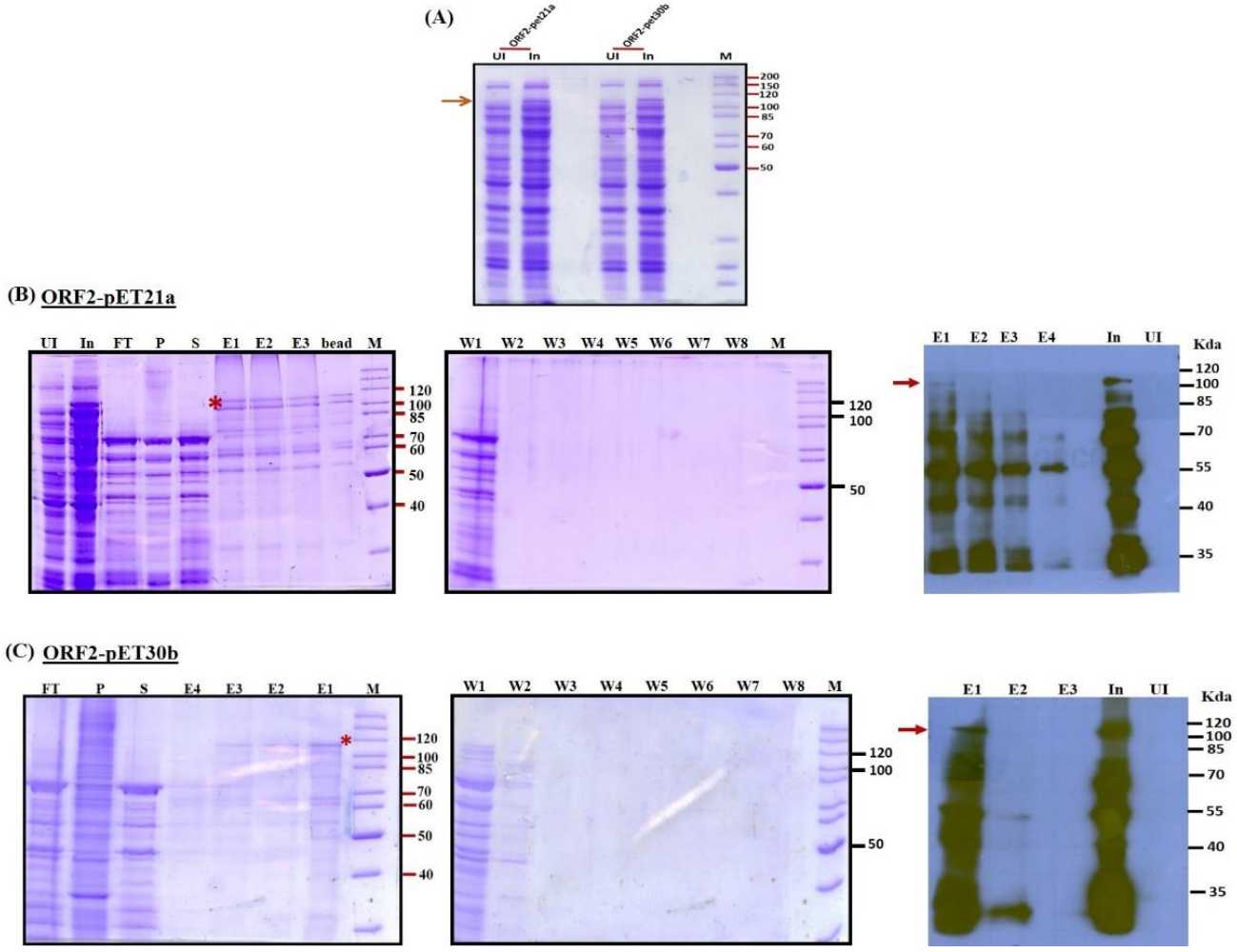


Figure 29: Expression and purification of ORF2-pET21a/pET30b: (A) Overexpression of ORF2-cloned in pET21a or pET30b vectors, in *E. coli* BL21 (RIL) strain, induced with 0.5mM IPTG/16ºC/6h. Lysates were prepared from Uninduced (UI) and IPTG-induced (In) cells. Unstained protein ladder was used as a size marker. (B and C) The supernatant (S) and pellet (P) fractions of cell lysate were analyzed. Purification of ORF2-pET21a was done using Ni-NTA Agarose resin. Flow through (FT); Eluted fractions (eluted with 150-250mM imidazole and 20mM DTT) (E1, E2, E3). Samples were resolved on 10% SDS-PAGE and proteins were visualized by Coomassie blue staining. Some protein remained stuck to the bead even after elution with 250mM imidazole. The asterisk marks the expected protein on the basis of size. Western blot analysis was performed to see the expression of recombinant proteins using Anti-His antibody. Arrow indicates the expected size band of interest. Semi dry transfer on PVDF membrane was performed for immunoblotting and the Chemiluminescence signal was detected with Millipore immobilin kit.

Expression using ORF2-pET30b and ORF2-pET21a vectors was further checked in SHuffle cells that are engineered *E. coli* K12. These cells constitutively express a copy of the disulfide bond isomerase which promotes the correction of mis-oxidized proteins and is recommended for cysteine-rich proteins (Lobstein *et al.*, 2016; Rosano *et al.*, 2014). However, we were not able to improve the

expression of ORF2p in SHuffle cells (fig.30A). Keeping in mind that codon biasness may be the reason of not getting the expression in SHuffle cells, ORF2 was cotransformed with Rosetta/RIL in SHuffle cells (Fig.30B). Separately SHuffle cells were also co-transformed by pRare plasmid (Fig.30C). pRare encodes tRNA genes for all of the "problematic" rarely used codons to enhance protein expression from target genes containing rare *E. coli* codons that would otherwise impede translation. Unfortunately, we could not get the expression in either cotransformed SHuffle cells.

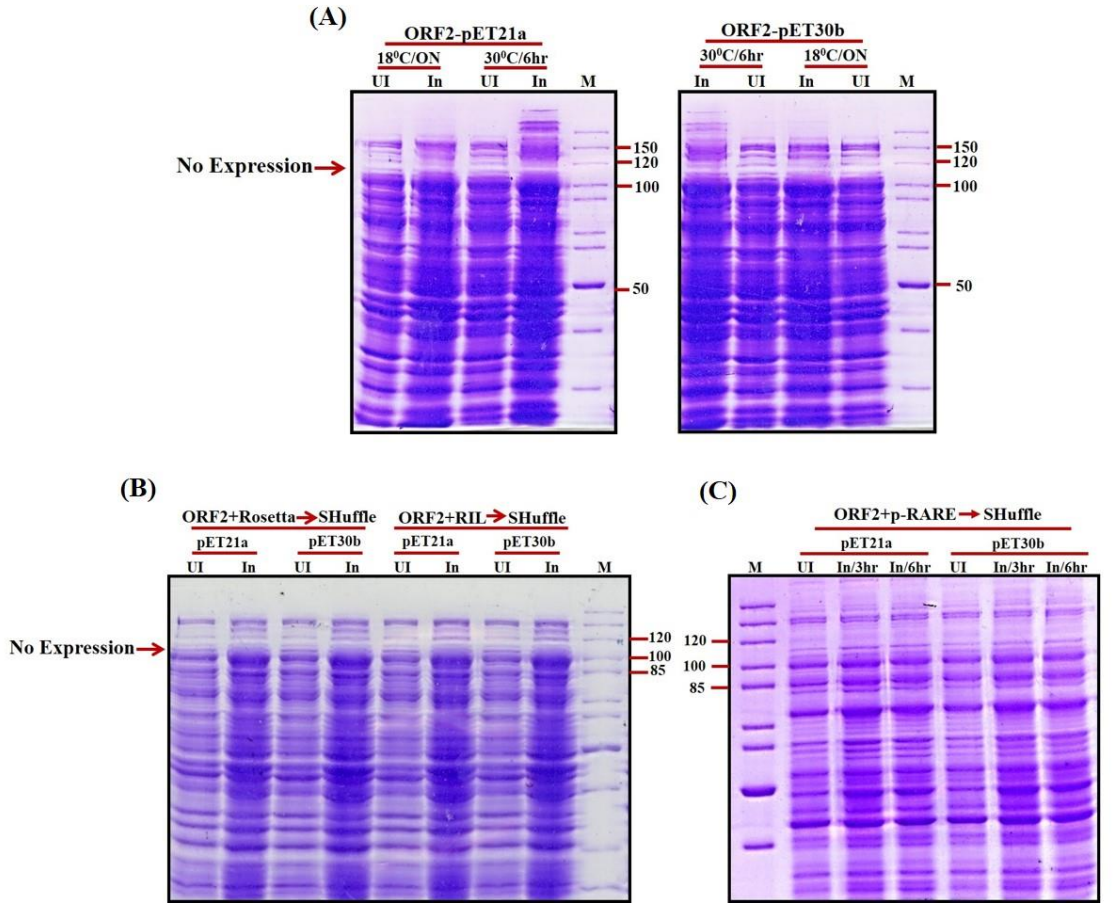


Figure 30: Overexpression of ORF2-pET21a/pET30b in Shuffle cells. (A) ORF2 cloned in pET21a and pET30b expression vectors were transformed into *E. coli* Shuffle cells followed by overexpression at 18°C/ON and 30°C/6h with 0.5mM IPTG. Uninduced (UI) and induced (In) lysates were prepared and resolved on 10% SDS-PAGE as mentioned in figure 29. (B) ORF2-pET21a/pET30b co-transformed with Rosetta and RIL separately into Shuffle cells and overexpressed at 30°C/6h with 0.5mM IPTG. (C) ORF2-pET21a/pET30b co-transformed with p-RARE into Shuffle cells and overexpressed at 30°C/3h, 6h with 0.5mM IPTG.

Further, we expressed ORF2 that was cloned in *BamH*I- *Xho*I site of pGEX-4T1 (GST tagged) expression vector, which gave us better result with less degradation. GST-tagged full-length ORF2p was expressed in *E. coli* (RIL) cells. The expression was checked with varying IPTG concentrations

and duration. Best expression was found with 0.5mM IPTG/6h/16ºC (Fig.31A). After induction, cells were pelleted and protein was affinity purified using glutathione sepharose as described in Materials and Methods. Purified protein showed the full-length band (137 kDa including GST tag) along with one major additional band as revealed by western blotting with anti- GST antibody (Sigma) (Fig.31B, C). The faster-migrating band might be degradation or processing product. Vector alone (pGEX-4T1) expressing the GST was also purified (Fig.32) and used as a control in subsequent assays to check the enzyme activity of reverse transcriptase and endonuclease.

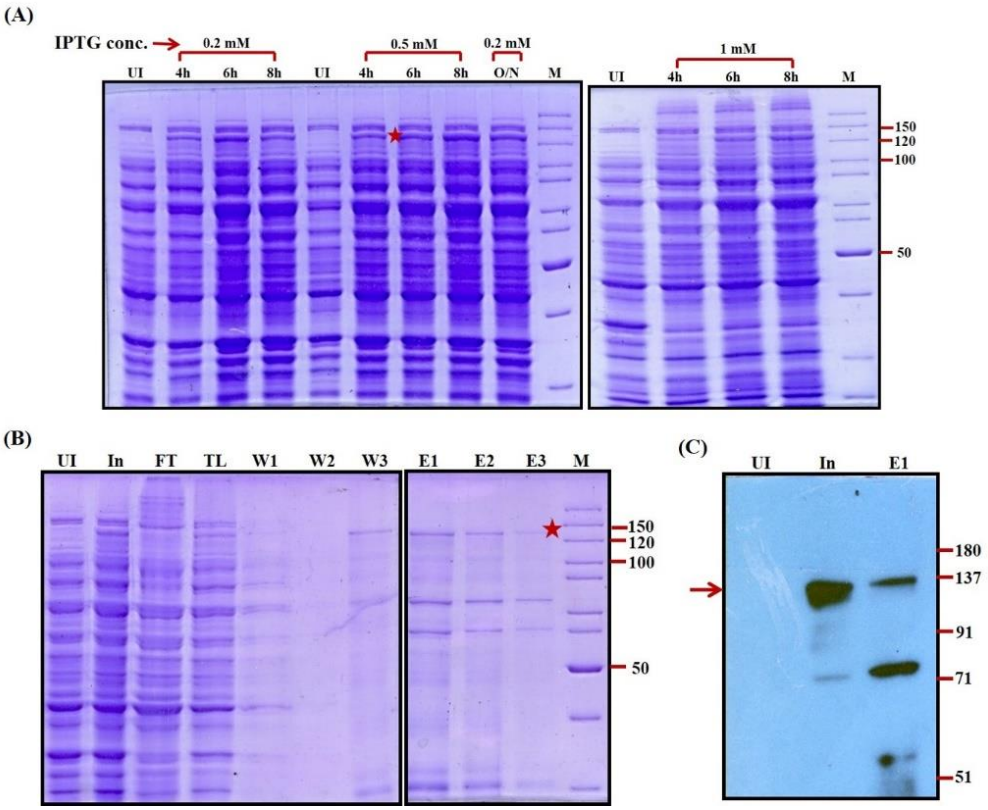


Figure 31: Overexpression and purification of ORF2-pGEX-4T1: (A) ORF2 cloned in pGEX-4T1 expression vector was overexpressed with indicated IPTG concentrations at 16ºC for different time duration. Induction with 0.5mM IPTG at 16ºC/6h showed better expression. (B) Purification showed protein degradation along with some nonspecific protein products. TL (total lysate); FT (flow through). Purified ORF2p was visualized by resolving on 10% SDS-PAGE and marked with an asterisk on the basis of expected size (137kda). (C) Western blot of purified ORF2p with anti-GST antibody showed some degradation of the protein.

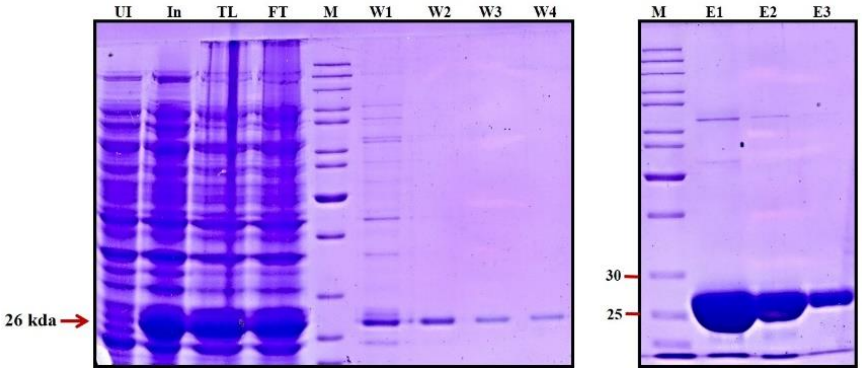


Figure 4: Purification of pGEX-4T1 vector alone. pGEX-4T1 vector was transformed into *E. coli* BL21 (RIL) cells and overexpressed at the same condition used for ORF2-pGEX-4T1. It showed strong expression along with abundant protein in different elution fractions.

3.6.2 Reverse transcriptase (RT) and Endonuclease (EN) activity with partially purified recombinant ORF2p

As ORF2p consists of RT and EN domains that are required for retrotransposition, we set up assays for RT and EN activity with the ORF2p from both pET30b and pET21a expression vector to check whether the protein was in active form. Though the protein was not very pure it did show RT activity as measured by RT-PCR, although not by direct RT assay. RT activity assay was performed with *in vitro*-transcribed 120nt EhSINE1 RNA (RNA 1) along with purified ORF2p; commercial revertaid RT was used as a positive control (Fig.33A). No cDNA product was detectable with purified ORF2p even after 120min, whereas we could get the expected product of 120nt with commercial RT at 60min (Fig.33B). Further, we tested RT activity by RT-PCR using cDNA synthesized from *in vitro*-transcribed 580nt EhSINE1 RNA (RNA 2) along with EhSINE1F and EhSINE1rev primers. The expected amplicon of 580bp was obtained in RT-PCR with purified ORF2p, showing that the protein does have detectable enzymatic activity (Fig.33C). Although the protein had visible RT-PCR activity, we could not see the EN (endonuclease) activity in protein expressed from either vector. pBS supercoiled plasmid DNA was used as the substrate for EN activity. Purified EN domain protein was used as the +ve control as it has been shown earlier to be enzymatically active (Mandal *et al*., 2004). Endonuclease converts the pBS plasmid DNA into an open circle followed by a linear form, both of which are seen in the +ve control (Fig.33D).

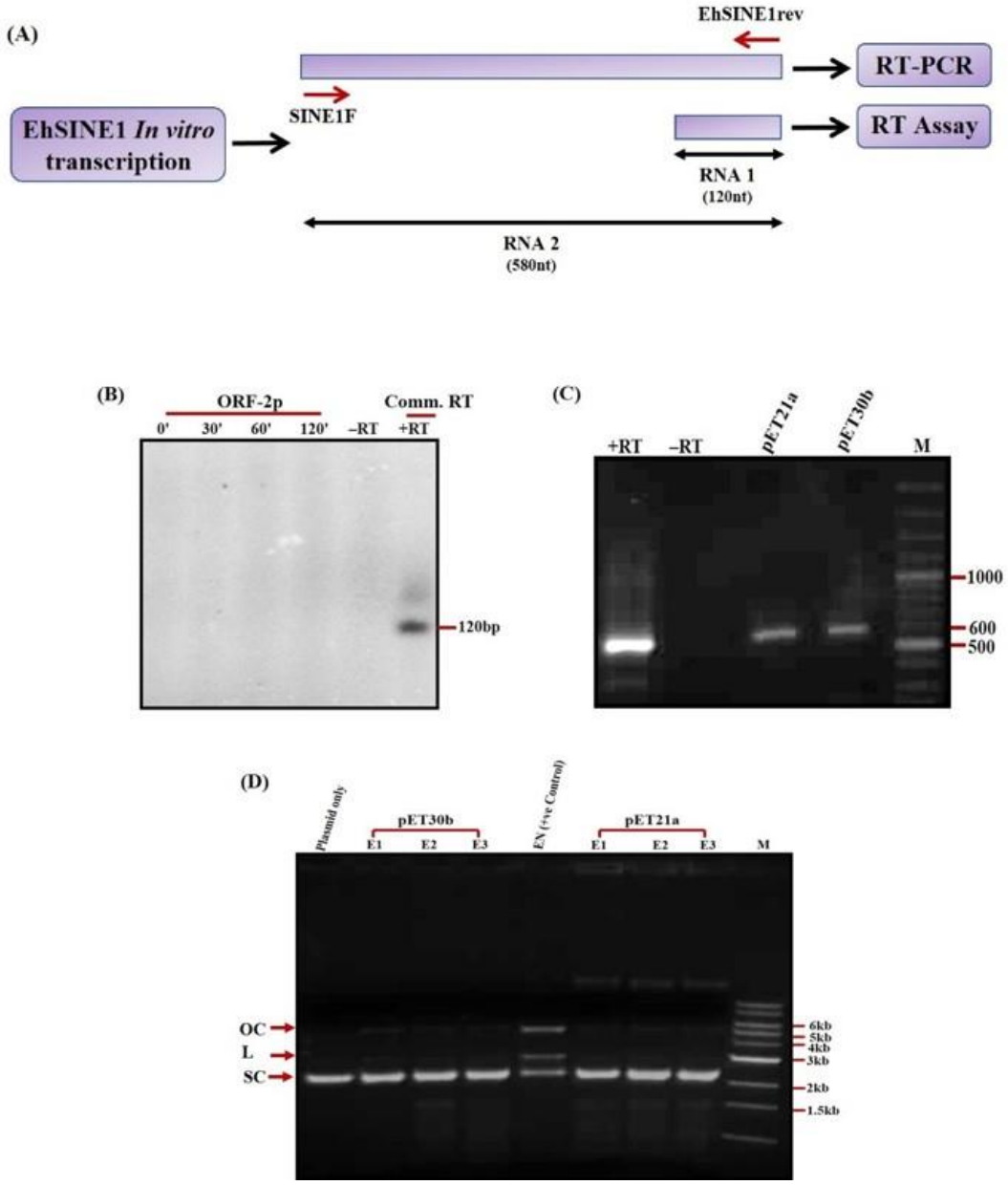


Figure 33: RT and EN assay with the recombinant ORF2-pET21a/pET30b polypeptides. (A) Schematic representation of EhSINE1 which was in vitro transcribed to obtain the template RNAs for RT assay. RNA 1 (120nt), RNA 2 (580nt) were reverse transcribed with EhSINE1rev primer. RT-PCR was done with SINE1F and EhSINE1rev primers. (B) RT assay with ORF2p at 42ºC. Commercial (Comm.) RT (Revertaid from Fermentas) for 60min at 42ºC used as the positive control and -RT without enzyme as the negative control. Final RT product was resolved on 10% denaturing PAGE containing 7M urea and autoradiographed. (C) Agarose gel (1.2%) picture of the RT-PCR assay with the purified recombinant ORF2p from both pET21a and pET30b expression vector along with 100bp marker. Commercial revertaid RT used as the positive control (+RT) whereas no enzyme is the negative control (-RT) (D) EN assay at 37ºC/60min with different elution fractions of the same recombinant ORF2p that was used for the RT. Purified EN domain protein was used as the positive control and products were checked on 0.8% agarose gel. Plasmid only- supercoiled pBS-plasmid. M, 1kb marker. [SC: supercoil; L: linear; OC: open circle]

Although there was no detectable EN activity from ORF2-pET30b/pET21a; we were able to find it at low level from ORF2-pGEX-4T1 (Fig.34A). After 120min of incubation, we could see the appearance of open circle form, although the linear form was not visible. Further, we checked RT activity in ORF2-pGEX-4T1. Like the ORF2-pET30b/pET21a, it also did not show activity in RT assay but showed significant activity after RT-PCR amplification (Fig.34B). We also confirmed the specificity of the reaction as it did not take place if either of the components (RNA, primer, and dNTPs) was excluded (Fig.34B). Extracts from vector alone (pGEX-4T1) were inactive, showing that the observed activity was not due to any host cell contaminant (Fig35).

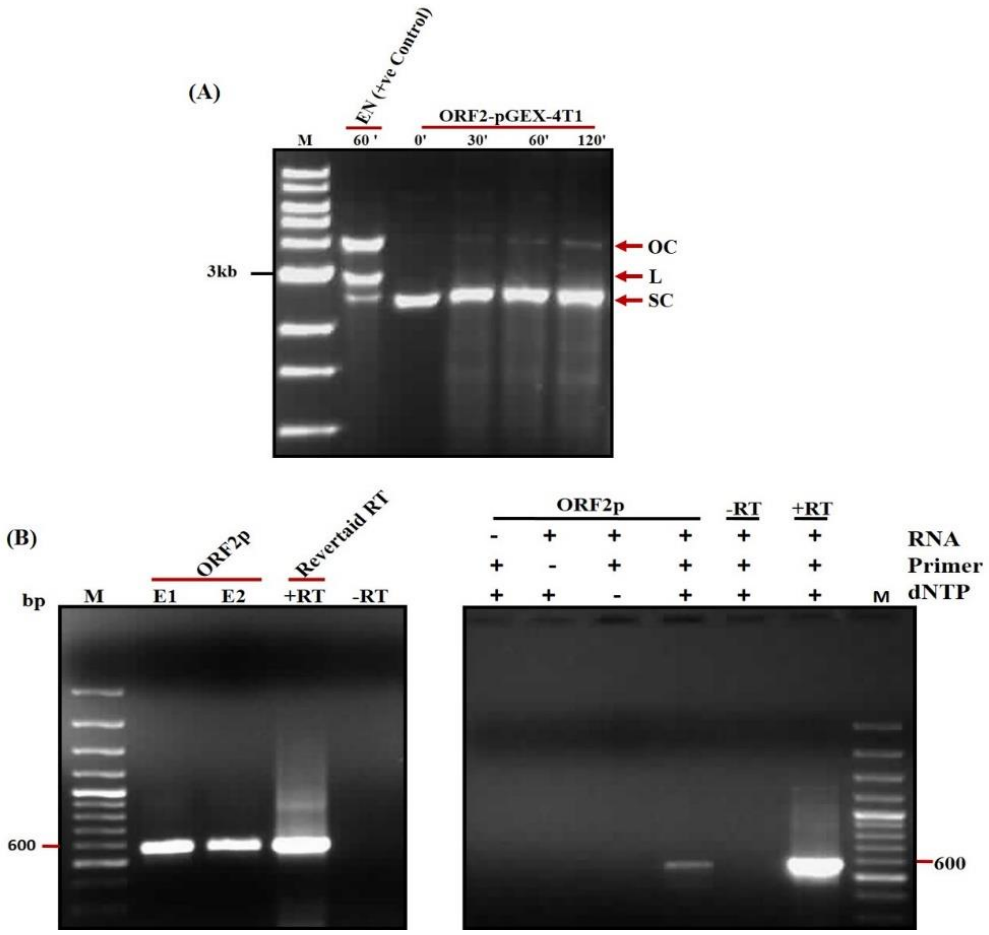


Figure 34: EN and RT assay with ORF2-pGEX-4T1. (A) Agarose gel (0.8%) picture of EN assay with ORF2p at 37ºC for 0, 30, 60 and 120min. ORF2p could convert some of the pBS DNA into OC form (B) RT-PCR assay with elution fractions E1 and E2 showed the expected size amplicon on 1.2% agarose gel. Commercial RT (Revertaid) used as positive control and no enzyme as a negative control.

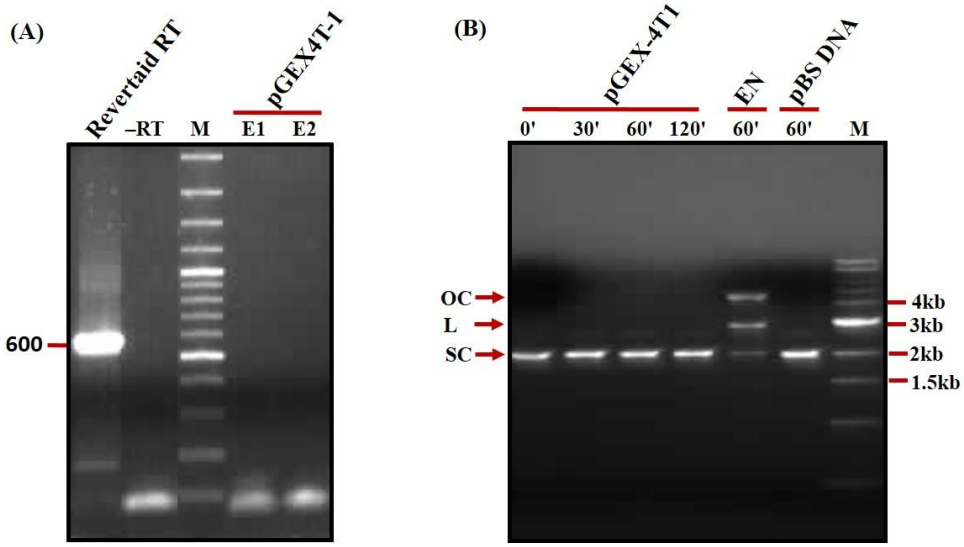


Figure 35: pGEX-4T1 vector control for RT and EN assay. Purified vector alone pGEX-4T1used for the RT and EN assay with the same conditions as used for ORF2p. (A) RT assay with pGEX-4T1 elution fraction 1 and 2. (B) EN assay with purified pGEX-4T1 at 37ºC for indicated times. Purified EN was used as the positive control.

From this study, we conclude that GST-tagged ORF2p is more active, probably because of additional stability provided by GST tag. However, the activity was insufficient to be used for demonstrating in vitro retrotransposition.

3.7 Cloning and Overexpression of ORF2p RT domain

3.7.1 Cloning of RT domain

It has been shown that EN domain alone expressed abundantly and was enzymatically very active, whereas in full-length ORF2p the EN activity was lost (Cost *et al*., 2002**)**. We wished to check whether this is the case with RT domain as well and whether the RT domain expressed separately might be more active. To check this, a fragment containing the RT domain (position 2659-4006 in EhLINE1) was cloned into pGEM-T Easy cloning vector, followed by sub cloning in pET30b expression vector. The cloned RT domain includes the putative active sites and the highly conserved YMDD motif that is required for RT activity (Larder *et al*., 1987; Harris *et al*., 1998) (fig.36).

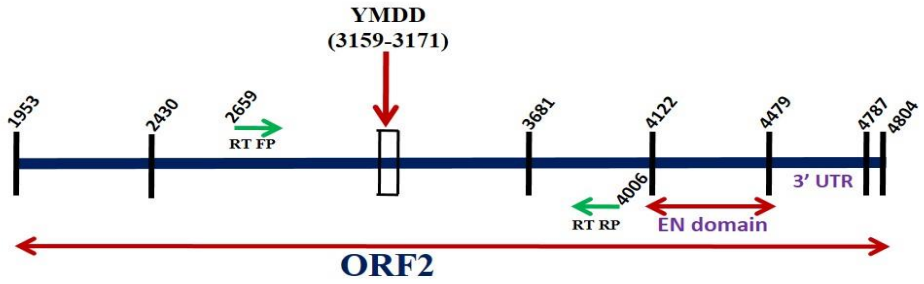


Figure 36: Cloning of RT domain. Schematic representation of primers (in green) to clone the RT domain

3.7.2 Expression and purification of RT domain

Recombinant RT domain (449aa) was expressed in Rosetta and RIL cells at 16^{o}C/6h with 0.2mM and 0.5mM IPTG. It showed the almost similar level of expression in both the cells at both IPTG concentrations. We found that most of the protein was going into inclusion body and very less amount was in the soluble fraction. We also tried Arctic cells at 10^{o}C/24h with 0.1 and 0.5mM imidazole but could not get good expression. Western blot showed that the expression was more in RIL cells compared to Rosetta cells (Fig.37A). The recombinant RT domain was purified from RIL cells using Ni-NTA agarose resin. Stringent washing with 20, 50 and 80mM imidazole was done to remove the contaminating proteins and eluted at 150 and 250mM imidazole concentration. Purified protein was then confirmed by western blotting using anti-His antibody. In addition to the band at the expected size of 53kDa, we also got a smaller band at ~27kDa, both in the induced cell lysate and in purified protein from column (Fig.37B).

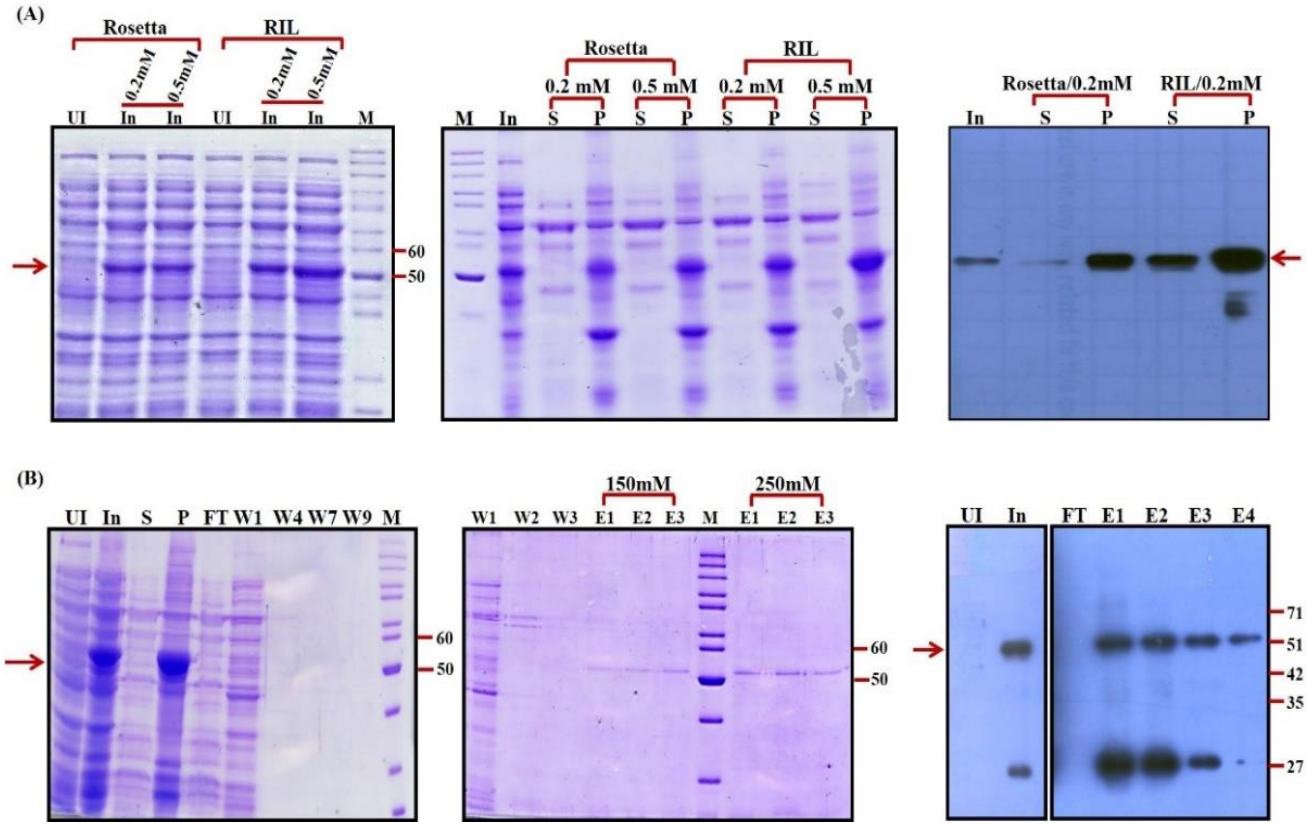


Figure 37: Expression and purification of RT domain in *E. coli* Rosetta and RIL cells: (A) RT domain cloned in pET30b expression vector was overexpressed in Rosetta and RIL cells at 16ºC/6h with 0.2mM and 0.5mM IPTG. It showed the almost similar level of expression at both concentrations in both the cells. Supernatant (S) and pellet (P) was checked after cell lysis and found that most of the protein was going to pellet. Western blot showed that some of the protein was also in the supernatant, which was used for purification. (B) RT-pET30b- RIL was purified by expressing at 16ºC/6h with 0.2mM IPTG conc. Cells were lysed and purified by affinity purification using Ni-NTA agarose resin. W1 to W9 is the different wash fractions. Elution was done in batch with low to high imidazole conc. E1, E2, and E3 are the elution fractions with 150mM and 250mM imidazole. Western blot analysis of purified protein along with uninduced (UI) and induced (In) cell lysate showed a faster migrating band in induced cells.

3.7.3 RT activity in the recombinant RT domain

RT assay with recombinant RT domain was done using the same approach used for ORF2p to check the protein activity. We found that the RT domain alone also had no visible activity in the RT reaction, and the activity was visible only after RT-PCR amplification as earlier shown for ORF2p (Fig.38). It was thus not useful to pursue in vitro retrotransposition assay with this protein.

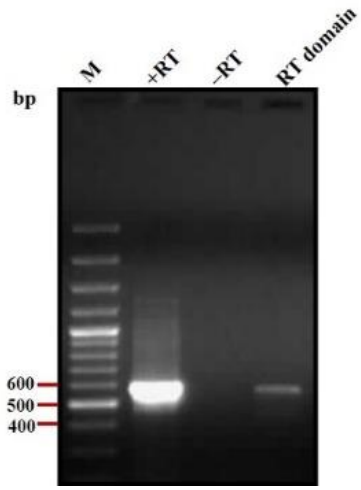


Figure 38: RT assay with recombinant RT domain: RT domain showed the activity similar to ORF2p. Commercial Revertaid RT used as the positive control (+RT) and no enzyme as the negative control (-RT). M, 100bp DNA ladder.

Discussion

4.1 Expression analysis of EhLINE1 and EhSINE1

Non-LTR retrotransposons are widespread in eukaryotic genomes and are of two sub-types, the autonomous long interspersed elements (LINEs) and the non-autonomous short interspersed elements (SINEs). LINEs possess either one or two ORF; the presence of one ORF in non-LTR retrotransposons has been considered to be an ancient feature (Kapitonov *et al*., 2009). Non- LTR elements represented in R2 group consist of one ORF encoding a single long polypeptide which contains all the activities required for retrotransposition. Although *E. histolytica* LINEs correspond to the R2 clade, some copies of EhLINE1 contain two ORFs due to presence of stop codon between ORF1 and ORF2, while some may also contain a single ORF in which stop codon is missing, as determined earlier in our lab through sequence analysis (Bakre *et al*., 2005). Studies reported in this thesis have been done with the consensus EhLINE1 copy containing two ORFs.

In *E. histolytica* some of the EhLINE1 copies are full-length but many are truncated either from 5'-end or 3'-end and a few copies are truncated from both of the ends (Bakre *et al*., 2005) which shows resemblance with the human L1 element. The majority of L1s are inactive due to mutations, truncations, and rearrangements. In L1, 5'-truncation can occur through transcription initiation near the downstream end of 5′-UTR (Alexandrova *et al*., 2012) due to the presence of multiple transcription start sites. On the other hand, 3'-truncation can be generated due to internal polyadenylation (Perepelitsa-Belancio *et al*., 2003). Genesis of truncated copies in EhLINEs has not been investigated. The 5'-UTR of EhLINE1 contains a promoter however, extensive deletion analysis has not been done to look for multiple start sites from multiple promoters. Although very few LINE copies are active; they contribute significantly to transcriptome diversity and gene regulation (Cordaux and Batzer, 2009). Truncated RNAs can be expected to mobilize provided that they encode the proteins required for transposition or these are provided in trans (Belancio *et al*., 2006; Moran *et al*., 1996). Mobilization in trans has been demonstrated for EhSINE1 (Yadav *et al*., 2012).

SINEs are highly abundant and known to affect gene expression in mammals (Kramerov *et al*., 2005). In *E. histolytica*, transcriptional silencing of Amoebapore which is believed to be involved in pathogenicity has been attributed to the adjacent EhSINE1 (Anbar *et al*., 2005; Mirelman *et al*., 2008) although other intergenic sequences may also be playing a role. Sequence analysis of the EhSINE1 copies in *E. histolytica* genome has been documented in detail (Huntley *et al*., 2010). However, studies on expression analysis of EhSINE1 as well as EhLINE1 have not been done so far. Hence, we undertook the expression analysis of these elements. We obtained sequence data by targeted sequencing of RT-PCR amplicons (Ion torrent) of EhLINE1 and EhSINE1 copies, and also

analyzed their transcription status from RNA-Seq data of total transcriptome (Illumina). The amplicon analysis was undertaken mainly for a cost-effective consideration as it would require less sequence coverage. With the reducing cost of NGS, it is now less relevant. In the amplicon data set, all 57 full-length EhLINE1 copies of the total 967 copies showed expression, while only 35% of them showed significant expression in the RNA-Seq data. This discrepancy could be because in RNA-Seq the copies with <10 read counts were not scored as expressed, while even low-level expression would be detectable by RT-PCR in the amplicon data. The same discrepancy was also observed in the truncated EhLINE1 copies; 30% of them were expressed in amplicon data while only 4% were expressed in RNA-Seq data. Expression analysis of EhSINE1 copies again showed discrepancy in the two data sets. There are 393 full-length EhSINE1 copies and 100 truncated copies. Of the full-length copies, 21% showed expression in amplicon data, while 38% showed expression in RNA-Seq data. None of the truncated copies showed expression in amplicon data while 6% of them were expressed in RNA-Seq data. This was reverse of the situation observed for EhLINE1 as more expressed copies were scored in RNA-Seq. We believe it could be related to the secondary structure of EhSINE1 due to which it may be inefficiently reverse transcribed in RT-PCR. Copies with low TPM value in RNA-Seq (2-5) were missed out from the amplicon data and the copies that showed expression in amplicon data had high TPM values (6-16). SINE1 copies with 2-5 TPM account for 29% that were missed from the amplicon data. Overall, the RNA-Seq data was much more informative as it gave a quantitative view of the expressed copies, and also provided a map of EhLINE1 with respect to the number of transcript reads coming from the entire length. Further analysis was done with RNA-Seq data.

4.1.1 Correlation of Expression data from RNA-Seq with Northern analysis

The distribution pattern of reads along the EhLINE1 showed that majority of reads aligned to the ORF2-RT domain, while a much small number aligned with ORF1 and ORF2-EN domain. No reads aligned to the spacer region between ORF1 and ORF2. In keeping with this, the northern data also showed no expression of the full-length transcript (4.8kb) (Fig.16). Probe from both ORFs hybridized with a short transcript of ~1.5kb (Yadav *et al*., 2012). Even in stressed conditions such as heat and oxidative stress (where LINE expression is known to be upregulated in other systems), we only got transcripts of ~1.5kb and the expression level was similar to normal growth conditions (Fig.16B). Our observation of a distinct transcript of 1.5kb from both the ORFs is different from the primate L1s and mouse L1 subfamilies in which a variety of transcripts of different sizes, including full-length has been shown. (Dudley *et al*., 1987; Belancio *et al*., 2006; Perepelitsa-Belancio *et al*., 2003; Nigumann *et al*., 2002; Speek, 2001). Both RNA-Seq and northern data of EhLINE1

expression showed that bulk of transcripts came from 1.5kb region of ORF2 encompassing the RT domain (nt position 2500-3800). ORF1 also gave rise to a 1.5kb transcript but of lower abundance as evident both from RNA-Seq and northern data. On the other hand, whereas RNA-Seq showed reads mapping to the EN domain of ORF2, we consistently did not find any signal in northern blots using EN probe. It is possible that transcripts from this region are very short in size and were missed in the northern analysis. The spacer region and part of the 5'-end of ORF2 (nt position 1511-2429) for which there were no reads in RNA-Seq, also gave no signal in northern blots. Hence, barring the EN domain, there was very good concordance between RNA-Seq and northern data for the entire EhLINE1. Transcripts lacking ORF1 and containing only ORF2 sequences have been reported in human L1, where they arise due to splicing (Belancio *et al*., 2006, 2010; Gasior *et al*., 2006). These spliced mRNAs are able to produce functional ORF2p which can introduce DNA double strand breaks (DSBs) or drive Alu retrotransposition. The functional significance of the ~ 1.5kb ORF2 transcript of EhLINE1 remains to be investigated. In future work, we will generate an antibody against the RT domain to check if this transcript is translated. Earlier work from our lab had shown the absence of full-length ORF2 polypeptide containing the RT and EN domains (using anti-EN antibody), and inability of *E. histolytica* cells to mobilize a SINE copy in the absence of ectopic expression of full-length ORF2 polypeptide (Yadav *et al*., 2012). In this work, we had not tried ectopic expression of EN domain alone. In case the constitutively present RT transcript is translated into active RT, it is possible that ectopic expression of EN domain alone may be sufficient for productive retrotransposition.

The LINE elements which are truncated and have lost the function may still contain promoter and polyadenylation sites that can regulate the transcription of neighboring genomic regions (Mourie *et al*., 2008; Wheelan *et al*., 2005). In EhLINE1, both ORF1 and ORF2 transcripts were found to be polyadenylated. Current data is insufficient to state that whether a single transcription event initiating from the promoter at the 5'-end of EhLINE1 gives rise to both ORF1 and ORF2 transcripts by processing of a precursor transcript, or whether ORF2 transcript originates from a second promoter located downstream of ORF1. If the latter is the case it would be a novel observation, since the ORF2 transcript in human L1 arises from alternative splicing of transcripts originating from the promoter in 5'-UTR. Though, we observed that there is no change in ORF1 transcript level during stress condition, we have earlier shown that ORF1 polypeptide levels drop significantly under the same conditions. Thus, translation of ORF1 mRNA is down regulated during heat stress. As ORF1p is a nucleic acid chaperone it may interact with specific RNAs in normal cells. It is possible that during stress some of these RNAs may get down regulated and not present

for interaction; thereby resulting in down regulation of ORF1 translation. Alternatively, ORF1p translation may be regulated independent of its interacting partners.

ORF1 and ORF2 3'-ends located by RT-PCR (Fig.20) also matched with RNA-Seq read distribution and is in agreement with the observed transcript size of 1.5kb. These transcripts could not be all arising from truncated copies since a significant number of full-length copies are transcribed, but no full- length transcripts are seen. Our data suggest extensive processing of EhLINE1 transcripts, which may involve internal termination, polyadenylation sites, or splicing events. Further work is required to understand the underlying mechanisms.

4.1.2 EhLINE1 promoter and transcript orientation

Mammalian L1s contain an internal promoter in their 5'-UTR which can synthesize retrotransposition competent transcripts (Swergold, 1990). On the other hand, R2 elements of *B. mori* do not use their own promoter and are thought to be co-transcribed with their host 28S rRNA (George *et al*., 1999). In EhLINE1 we have identified internal promoter at the 5'-end of ORF1, between150 to 200bp. The~ 1.5kb ORF2 transcript could well arise from this promoter, or there could be a second internal promoter upstream of ORF2. Our data so for did not show evidence for the latter, but further confirmation is required. Interestingly, we found antisense transcripts of ~ 1.5kb from ORF2 (RT) region, although the same were not found in ORF1 (Fig.18). The sense/antisense transcripts of ORF2 could form dsRNAs, which are known to be produced by retrotransposons and processed into dsRNAs from bidirectional transcripts (Watanabe *et al*., 2008). dsRNAs have been shown to have role in epigenetic regulation as well as mRNA stability (Tang *et al*., 2001; Yang *et al*., 2000). The antisense transcript of EhLINE1 RT region may be generated from an antisense promoter in 3'-region of ORF2 or it may be the result of read-through transcription from downstream gene, which needs to be explored. The well-characterized promoter of human L1 element located in the 5'-UTR contains both a 5'-sense promoter (Swergold, 1990) along with a downstream antisense promoter (ASP) between nt 400-600 of 5'-UTR (Speek, 2001). Due to the genome wide distribution of LINEs, the active ASP of human L1 can give rise to chimeric transcripts from a large number of neighbouring genes and are estimated to affect as many as 4% of all human genes (Criscione *et al*., 2016). Such divergent transcription has been shown for long noncoding RNA (lncRNA) in humans (>60%) which originated at promoters of protein coding genes (Sigova *et al*., 2013). Non-coding RNAs are known to regulate the epigenome by antisense transcription and are associated with chromatin modifications (Cruickshanks *et al*., 2013). It is possible that the ORF2-RT antisense transcripts of EhLINE1 may have a regulatory role to suppress

retrotransposition. Further work will involve identification of the promoter giving rise to these transcripts, and possible regulatory role of antisense transcription in EhLINE1.

4.2 Methylation status of LINE1 and heat shock protein gene (HSP70) in *E. histolytica*

DNA methylation is essential for normal development (Okano *et al*, 1999) and is uniquely achieved in all cell types (Bird, 2002; Reik, 2007; Ziller *et al*, 2013). Cytosine DNA methylation at promoter regions is a common mode of retrotransposon silencing in a variety of organisms (Law, 2010). Earlier studies in *E. histolytica* have indicated the possibility of methylation of EhLINE sequences since antibodies against the *E. histolytica*-methylated LINE binding protein (EhMLBP) interacted with EhLINE sequences in vivo (as shown by chromatin immunoprecipitation) (Lavi, 2006). The protein also bound to another highly repetitive DNA- the rDNA of *E. histolytica*, suggesting that it could have a role in modulating the expression of highly repetitive DNA. However, direct demonstration of promoter methylation and transcription attenuation of these sequences has not been done. Hence, we started looking for the level of cytosine methylation at the EhLINE1 promoter and its correlation with transcriptional repression.

In our analysis, we found that bisulfite sequencing of a 200bp region at the 5'-end of an expressed and silent EhLINE1 copy showed complete lack of cytosine DNA methylation in both copies. To confirm that this was not due to a technical problem we showed that the Hsp70 promoter was fully methylated in our cells, as previously reported (Fisher *et al*., 2006). Since EhLINE1 is present in 967 copies it is possible that some of these copies may be methylated, and the two copies analysed by us were exceptions. We used the strategy of single nucleotide incorporation opposite cytosine in bisulfite treated DNA to check the methylation status of selected cytosines in a larger subset of EhLINE1 (62.5% and 78.5% of the 5'-intact copies contained CpG or non CpG cytosines, respectively). Again, we did not find any cytosine methylation, whereas all the cytosines in Hsp70 were scored as methylated by this method also. Our data shows that EhLINE1 promoter sequences are almost devoid of cytosine DNA methylation. It is possible that the transcriptional status of EhLINEs may be regulated by other mechanisms which remain to be explored. For example, histone methylation instead of DNA methylation might suppress transcription of these elements, as reported for some mammalian SINE sequences (Elbarbary *et al*, 2016). The study of Fisher *et al*. (2006) looked at phenotypic changes in cells overexpressing Ehmeth, a methyltransferase of the Dnmt2 family. These cells showed pleiotropic changes (multinucleation, resistance to oxidative stress), and the transcription of HSP70 gene was upregulated. Since this gene is fully methylated even in normal cells, its methylation status is not expected to change in the Ehmeth-overexpressed cells, and the observed upregulation could be an indirect effect. We provide direct evidence that the

E. histolytica HSP70 promoter DNA remains methylated during heat shock when the gene is actively transcribed. Thus, cytosine methylation is not a repressive mark for HSP70 as well. Results showed that DNA cytosine methylation of promoter, which is a common mechanism of transposon silencing in a variety of organisms, is unlikely to modulate the transcription of EhLINE1. The global mechanism responsible for transcriptional regulation of the large number of EhLINE1 copies in *E. histolytica* remains to be discovered.

4.3 Expression and purification of recombinant EhORF2p in *E. coli*

A functional full-length L1 element contains two long open reading frames, ORF1 and ORF2, required for retrotransposition (Feng *et al*.,1996; Moran *et al*., 1996). ORF1 encodes RNA binding protein which shows nucleic acid chaperone activity in vitro (Kolosha *et al*., 1997, 2003; Martin *et al*., 2001) and ORF2 encodes a protein with two functional domains; reverse transcriptase (RT) and endonuclease (EN) that are required for the RNP formation and retrotransposition (Moran *et al*., 1996). Similarly, EhLINE1 also contains two functional proteins, ORF1 and ORF2. During the process of retrotransposition, ORF2p nicks the target DNA and it uses the so generated 3'-OH to prime the reverse transcription of L1 RNA (Luan *et al*., 1993; Cost *et al*., 2002). The ORF2 protein has not been detected in human or mouse cells. However, cloned, transpositionally-active L1*Hs* elements have been used to demonstrate RT activity in transfected cell lines (Sassaman *et al*., 1997; Holmes *et al*., 1994; Dombroski *et al*., 1994; Mathias *et al*., 1991). Similarly, in *E. histolytica*, we could not get expression of ORF2 protein. However, in our case ORF1 showed endogenous expression. Moreover, it was very difficult to purify and detect full length recombinant EhORF2p from *E. coli* in comparison to other *E. histolytica* proteins studied in our lab. Detection of full length recombinant ORF2p has also been a major challenge in humans and rats as antibodies against ORF2p showed very low signal strength (Ergün *et al*., 2004; Goodier *et al*., 2004; Kirilyuk *et al*., 2008; Doucet *et al*., 2010). Poor expression of ORF2 has been linked to inadequate transcriptional elongation (Han *et al*., 2004). EhLINE1 ORF2p sequence analysis showed many codons for some amino acids (e.g. Arg, Ile and leu) which are less frequently present in *E. coli*. These codons were dispersed throughout the ORF2 sequence, which eliminates the possibility of sequence correction. To avoid this problem, we used codon-plus *E. coli* strains [Rosetta and BL21 (RIL)] for the expression of ORF2p. We tried various host vector combinations and expression conditions to achieve good amount of purified protein in its active form (summarized in table 6). We were able to express full length recombinant ORF2p to some level in BL21 (RIL) expression cells. To achieve acceptable level of active protein, ORF2 was tagged with 6X His sequences at both N-terminal and C-terminal along with N-terminal GST-fusion tag separately. ORF2p

expression and activity was better in GST-fusion protein in comparison to 6X His tagged protein. It has been earlier observed that short epitope tags are less efficacious to successfully detect ORF2p (Cost *et al*., 2002). GST-fusion tag has been used to enhance solubility and stabilization of the partner protein (Malhotra *et al*., 2009). Several mechanisms of action for GST fusion tags has been hypothesized but exact mechanism remains unclear (Nallamsetty and Waugh, 2007; Butt *et al*., 2005). GST tag is observed to protect its partner protein from proteolytic degradation (Kaplan *et al*., 1997; Hu *et al*., 2008; Young *et al*., 2012). Though we were able to purify ORF2p but it was not in its pure form in either tag, we got some non-specific proteins and some degradation products.

4.3.1 Reverse transcriptase and Endonuclease activity in recombinant ORF2p

The RT domain contains seven evolutionarily conserved RT motifs (Lingner *et al*., 1997; Nakamura *et al*., 1997; Cote and Roth, 2008). Of these, motifs RT3- RT7 lie in the active site for polymerization. The putative active site contains 3 aspartate residues; 1 residue located in RT3 and 2 in RT5 motif (Kohlstaedt *et al*., 1992). To check the RTp activity, RT domain (position 2659-4006 in EhLINE1, that contains motifs 3 to 7) was cloned into pET30b expression vector. The cloned RT domain includes the putative active site and motifs that are required for the RT activity (Larder *et al*., 1987; Harris *et al*., 1998). In *E. histolytica*, RT domain contains methionine residue at the place of "X" in RT5 "YXDD" motif which makes it "YMDD" as in lentiviruses, and such replacement has been thought to be responsible for the low fidelity of RTs (Kaushik *et al*., 2000; Poch *et al*., 1989).

Human L1 ORF2p RT is a highly processive polymerase (Piskareva *et al*, 2006). Both the domains RT and EN interact with their own L1 RNA in a *cis* preference manner (Wei *et al*, 2001) and forms ribonucleic acid particles (RNP) that are the intermediates of retrotransposition (Kulpa *et al*, 2006; Goodier *et al*, 2007; Doucet *et al*, 2010). We found that unlike the L1 ORF2p, EhLINE1 ORF2p is not highly processive. In human L1, RT activity could be seen in the form of cDNA by primer extension (Piskareva and Schmatchenko, 2006) whereas in EhLINE1 we could not see RT activity by primer extension in any of the tagged recombinant proteins; it could be visualized only after PCR amplification of the cDNA. Both the activities RT and EN were checked with N-ter, C-ter His tagged and GST-fusion tagged recombinant proteins. His tagged protein showed very low RT activity that can be seen after PCR amplification. Although it showed weak RT activity, we could not see any EN activity with the same protein. Active EN first nicks the supercoiled plasmid into open circle form followed by linear form. Though we could not get EN activity with His tagged protein, however, overexpressed GST tagged protein showed weak EN activity. We have earlier shown robust EN activity using the recombinant EN domain alone (Mandal *et al*., 2004). Repressed

EN activity in the ORF2 polypeptide may be due to change in the conformation of EN domain in the context of full-length ORF2 protein that may render it unable to efficiently nick the DNA, as shown by Cost *et al* (Cost *et al*, 2002).

4.3.2 Reverse transcriptase activity in recombinant EhRT domain

It has been shown with L1 that EN domain alone expressed abundantly and was enzymatically very active whereas in full-length ORF2p the EN activity was lost (Cost *et al*., 2002). To see whether the RT domain alone might be more active in our case, it was cloned separately into pET-30b cloning vector and overexpressed in Rosetta/RIL expression cell. Similar to full-length ORF2p, RT domain showed better expression in RIL cells compared to Rosetta. Therefore, further expression was done in RIL cells for the purification using Ni-NTA resin. It showed some degradation during induction but we were able to get sufficient protein to check the activity. RT activity was monitored with the recombinant RT domain as earlier done with full-length ORF2p. Similar to ORF2p, RT domain could not show RT activity in the form of cDNA, it could be visualized only after PCR amplification of cDNA product. It is necessary to have active RT protein along with EN protein to be able to demonstrate TPRT reaction *in vitro*. Since we do have enzymatically active EN recombinant polypeptide, more standardization needs to be done to get the highly active RT domain that can be used with EN domain to see the TPRT process *in vitro*.

Summary

Entamoeba histolytica is a microaerophilic protozoan parasite that lives in the human intestine and is the causative agent of amoebiasis. It is prevalent in unhygienic living conditions and is endemic in developing countries. Approximately 50 million people get infected with the parasite worldwide, causing 40 thousand to 1 lakh death per year.

Several families of transposable elements have been identified in the genome of *E. histolytica*. The most abundant of these is the non-Long Terminal Repeat retrotransposons, which occupy 11.2% of the genome. These consist of the long- and short-interspersed nuclear elements (LINEs and SINEs). *E. histolytica* genome contains three families each of LINEs (EhLINE1, 2 and 3) and SINEs (EhSINE1, 2 and 3). LINEs are autonomous elements encoding reverse transcriptase (RT) and endonuclease (EN) activities whereas SINEs are nonautonomous elements which use LINE machinery for their retrotransposition. EhLINE1 (4.8kb) and EhSINE1 (550bp) are the most abundant LINEs and SINEs in *E. histolytica* genome respectively. There are 967 copies of EhLINE1, of which 57 are full-length. Despite the presence of multiple EhLINE1 copies, none appear to be active due to accumulated mutations and truncations. The full-length EhLINE1 (4.8kb) element with complete ORFs was reconstructed in our lab previously. EhLINE1 encodes two non-overlapping ORFs (ORF1 and ORF2). ORF2p contains a centrally localized RT and a C-terminal EN domain, while ORF1p has nucleic acid–binding properties. ORF1p, ORF2p and LINE1 mRNA associate to form a ribonucleoprotein in human (L1 element), which is required for retrotransposition. It was previously shown in our lab that EhLINE1 ORF1p was constitutively expressed in *E. histolytica*, and active retrotransposition could be detected in these cells upon ectopic expression of ORF2p. However, the expression status of the multiple EhLINE1 copies, and of the two ORFs was not understood in detail.

In the present study was have undertaken the expression analysis of EhLINE1 and EhSINE1 in *E. histolytica* with a view to determine the expression status of the multiple genomic copies of these elements. Expression status was determined by two approaches- (1) targeted sequencing of RT-PCR amplicons from EhLINE1/SINE1, (2) total RNA-Seq data. The results were further validated and interpreted by northern blot analysis and other techniques. Since many copies were transcriptionally silent we checked the involvement of cytosine DNA methylation at EhLINE1 promoter region to see if it had a role in silencing. In order to study retrotransposition *in vitro,* we over expressed recombinant ORF2 protein in a bacterial system and determined enzymatic activities of the RT and EN domains which are required for retrotransposition.

Main findings of the present study are summarized below:

1. All EhLINE1 and EhSINE1 copies available in the database were categorized as full-length or truncated on the basis of alignment using the multiple sequence alignment tool MAFFT.

2. Expression status of EhLINE1/EhSINE1 copies was determined by targeted sequencing of RT-PCR amplicons, and total RNA-Seq data. Further analysis was limited to RNA-Seq data as it was more quantitative. A larger fraction of full-length copies were transcriptionally active than truncated copies (72% versus 28%). In EhLINE1 the bulk of reads mapped to a ~1.3kb region in RT domain of ORF2, which was flanked on both sides by regions with zero reads. A smaller number of reads mapped to ORF1 and to the EN domain of ORF2.
3. Validation of transcriptionally active and silent copies scored by RNA-Seq was done experimentally by RT-PCR, followed by cloning and sequencing the amplicons.
4. Distribution of reads within EhLINE1 revealed by transcriptome data correlated very well with northern blot analysis. Probes from the ~1.3kb transcriptionally active region of ORF2 RT gave strong signals in the northern analysis, while no signals were obtained from the flanking regions. ORF1 probe gave weaker signals. However, no signal was obtained with EN probe although some reads were scored from this region in RNA-Seq. It is possible that these transcripts may be short and heterogeneous in size hence, undetectable by northern analysis.
5. Strand specific probes revealed the presence of both sense and antisense transcripts from ORF2-RT region, while only sense transcripts were found for ORF1.
6. RT-PCR with oligodT primer showed 3'-polyadenylation of ORF1 and ORF2 transcripts.
7. The 3'-ends of ORF1 and ORF2 transcripts were located by RT-PCR with oligodT primer and several upstream primers. From the amplicon sizes the approximate 3'-ends were determined, which correlated with the transcriptome data.
8. Mapping of 5'-end of ORF2 by primer extension correlated very well with the transcriptome data.
9. Luciferase reporter assay showed the absence of an internal promoter in the spacer region between ORF1 and ORF2. However, these data require confirmation.
10. Cytosine DNA methylation at promoter regions of selected transcriptionally active or silent EhLINE1 copies was determined using bisulfite treatment followed by Sanger sequencing. No cytosine methylation was detected in either of the copies. Further, instead of comparing only two copies, we set up an assay to detect cytosine methylation of a few selected conserved cytosines in the 500bp region promoter region of a larger subset of EhLINE1 copies, using bisulfite treatment and single nucleotide incorporation with end labeled primer. Again, we found no methylation in either of the selected CpG sites. Thus, cytosine DNA methylation at promoter region does not seem to be involved

in transcriptional regulation of these elements. As a control, we looked at EhHsp70 gene in which cytosines in the promoter region are known to be methylated at normal growth temperature when the gene is silent. We showed that this gene continued to be methylated during heat stress when its transcription increased many fold. Thus, cytosine DNA methylation does not seem to serve as a transcriptional repressor mark in *E. histolytica.*

11. RT and EN assays were performed with purified full-length ORF2 protein, which showed very weak activity not sufficient for *in vitro* target primed reverse transcription assay. RT activity with purified RT domain was also too low. This needs to be optimized.

In conclusion, we have for the first time provided a detailed picture of transcription status of EhLINE1/SINE1 copies in *E. histolytica*. The bulk of EhLINE1 transcripts mapped to the RT domain as seen both by RNA-Seq and northern data. This study throws open a plethora of questions regarding the genesis of these transcripts, whether from an independent promoter or through processing of precursor transcript; the role of antisense RT transcripts; the translational efficiency of the sense RT transcripts; and mechanisms of transcriptional silencing other than cytosine DNA methylation.

Bibliography

- Adey N.B., Tollefsbol T.O., Sparks A.B., Edgell M.H., Hutchison C.A. Molecular resurrection of an extinct ancestral promoter for mouse L1**.** Proc Natl Acad Sci USA. 1994; 91:1569-1573.

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

- Alexandrova E.A., Olovnikov I.A., Malakhova G.V., Zabolotneva A.A., Suntsova M.V., Dmitriev S.E., Buzdin A.A. Sense transcripts originated from an internal part of the human retrotransposon LINE-1 5′ UTR. Gene. 2012; 511:46-53.

- Ambros V. The functions of animal microRNAs. Nature. 2004; 431:350-355.

- Anbar M., Bracha R., Nuchamowitz Y., Li Y., Florentin A., Mirelman D. Involvement of a Short Interspersed Element in Epigenetic Transcriptional Silencing of the Amoebapore Gene in *Entamoeba histolytica***.** Eukaryot Cell. 2005; 4:1775-1784.

- Anzai, T., Takahashi, H., and Fujiwara H. Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. Mol Cell Biol. 2001; 21:100-108.

- Athanikar J.N., Badge R.M., Moran J.V.A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. 2004; 32:3846-3855.

- Aurrecoechea C., Barreto A., Brestelli J., *et al*. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. Nucleic Acids Res. 2011; 39: D612-9.

- Aurrecoechea C., Brestelli J., Brunk BP., *et al*. EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res. 2010; 38: D415-9.

- Aurrecoechea C., Heiges M., Wang H., Wang Z., *et al.* ApiDB: integrated resources for the apicomplexan bioinformatics resource center. Nucleic Acids Res. 2007; 35: D427-30.

- Bakre A.A., Rawal K., Ramaswamy R., Bhattacharya A., Bhattacharya S. The LINEs and SINEs of *Entamoeba histolytica*: comparative analysis and genomic distribution. Exp Parasitol. 2005; 110(3):207-13.

- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116, 281–297.

- Beck P., Dingermann T., and Winckler, T. Transfer RNA gene-targeted retrotransposition of *Dictyostelium* TRE5-A into a chromosomal UMP synthase gene trap. J Mol Biol. 2002; 318:273-285.

- Belancio V.P., Hedges D.J., Deininger P. LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res. 2006; 34:1512-1521.

- Belancio V.P., Roy-Engel A.M., Pochampally R.R., Deininger P. Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res. 2010; 38:3909-3922.

- Belancio V.P., Whelton M., Deininger P. Requirements for polyadenylation at the 3' end of LINE-1 elements. Gene. 2007; 390(1-2):98-107.

- Bhattacharya S., Bakre A., and Bhattacharya A. Mobile genetic elements in protozoan parasites. J. Genet. 2002; 81:73-86.

- Bhattacharya S., Bhattacharya A., Diamond L.S. Comparison of repeated DNA from strains of *Entamoeba histolytica* and other *Entamoeba*. Mol Biochem Parasitol. 1988; 27(2-3):257-62.

- Bibillo A., and Eickbush, T.H. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. J Mol Biol. 2002; 316:459-473.

- Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002; 16:6-21.

- Blesa D., Martinez-Sebastian M.J. bilbo, a non-LTR retrotransposon of *Drosophila* subobscura: a clue to the evolution of LINE-like elements in *Drosophila*. Mol Biol Evol 1997; 14:1145-1153.

- Boeke J.D. LINEs and Alus--the polyA connection. Nat. Genet. 1997; 16**:**6-7.

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

- Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. Ty elements transpose through an RNA intermediate. Cell. 1985; 40:491-500.

- Bolger A.M., Lohse M., Usadel B. Trimmomatic: flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

- Bonchev G, Parisod C. Transposable elements and microevolutionary changes in natural populations. Mol Ecol Resour,. 2013; 13(5):765-75.

- Bourc'his D., and Bestor T.H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature. 2004; 431:96–99.

- Brinkac L.M., Davidsen T., Beck E., *et al*. Pathema: a clade-specific bioinformatics resource center for pathogen research. Nucleic Acids Res. 2010; 38: D408-14

- Brouha B., Schustak J., Badge R.M., Lutz-Prigge S., Farley A.H., Moran J.V., Kazazian H.H. Jr. Hot L1s account for the bulk of retrotransposition in the human population**.** *Proc Natl Acad Sci USA.* 2003; 100:5280-5285.

- Brumpt. *Entamoeba dispar* n sp amibe a kystes quadrinuclees, parasite de l'homme. Bull acad méd (Paris). 1925; 94:943–952.

- Brunmeir R., Lagger S., Simboeck et al. Epigenetic regulation of a murine retrotransposon by a dual histone modifcation mark. PLoS Genet. 2010; 6: e1000927

- Burke, W.D., Muller, F., and Eickbush, T.H. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. Nucleic Acids Res. 1995; 23:4628-4634.

- Butt T.R., Edavettal S.C., Hall J.P., Mattern M.R. SUMO fusion technology for difficult-toexpress proteins. Protein Expr Purif. 2005; 43(1):1-9.

- Cappello J., Cohen S.M., Lodish H.F. *Dictyostelium* transposable element DIRS-1 preferentially inserts intoDIRS-1 sequences.MolCellBiol. 1984; 4:2207-2213.

- Carrieri, C. *et al.* Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. Nature. 2012; 491:454-457.

- Cedar H., Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 2009; 10:295-304.

- Chaboissier, M.C., Finnegan, D., and Bucheton, A. Retrotransposition of the I factor, a non-long terminal repeat retrotransposon of *Drosophila*, generates tandem repeats at the 3' end. Nucleic Acids Res. 2000; 28:2467-2472.

- Christensen S., Pont-Kingdon G., Carroll D. Comparative studies of the endonucleases from two related Xenopus laevis retrotransposons, Tx1L and Tx2L: target site specificity and evolutionary implications. Genetica. 2000; 110(3):245-56.

- Christensen, S. M., Ye, J., and Eickbush, T.H. RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. Proc Natl Acad Sci U S A. 2006; 103:17602-17607.

- Clark C.G., Ali I.K., Zaki M., Loftus B.J., Hall N. Unique organization of tRNA genes in *Entamoeba histolytica*. Mol Biochem Parasitol. 2006; 146(1):24-9.

- Clark C.G., Alsmark U.C., Tazreiter M., *et al.* Structure and content of the *Entamoeba histolytica* genome. Adv Parasitol. 2007; 65:51-190.

- Clark, C.G., and Diamond, L.S. Ribosomal RNA genes of 'pathogenic' and 'nonpathogenic' *Entamoeba histolytica* are distinct. Mol Biochem Parasitol. 1991; 49:297-302.

- Clark, C.G. Cryptic genetic variation in parasitic protozoa. J Med Microbiol. 2000; 49:489-491.

- Contursi C., Minchiotti G., Di Nocera P.P. Identification of sequences which regulate the expression of *Drosophila melanogaster* Doc elements**.** *J Biol Chem.* 1995; 270:26570-26576.

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

- Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. Human L1 element target-primed reverse transcription in vitro. Embo J. 2002; 21:5899-5910.

- Cote M.L. and Roth M.J. Murine leukemia virus reverse transcriptase: structural comparison with HIV-1 reverse transcriptase. Virus Res. 2008; 134:186–202

- Craig N.L. Target site selection in transposition. Annu Rev Biochem. 1997; 66:437-474.

- Craig N. An introduction, In Mobile DNA II, N. Craig, R. Craigie, M. Gellert, and A. Lambowitz, eds. (American Society for Microbiology: Washington DC). 2002; pp. 3-10.

- Criscione S.W., Theodosakis N., Micevic G., Cornish T.C., Burns K.H., Neretti N., Rodić N. Genome-wide characterization of human L1 antisense promoter-driven transcripts. BMC Genomics. 2016;17:463

- Crowther P., Doherty J., Linsenmeyer M., Williamson M., Woodcock D. Revised genomic consensus for the hypermethylated CpG island region of the human L1 transposon and integration sites of full-length L1 elements from recombinant clones made using methylation-tolerant host strains. Nucleic Acids Res. 1991; 19(9):2395–401.

- Cruickshanks H.A., Vafadar-Isfahani N., Dunican D.S., Lee A., Sproul D., Lund J.N., Meehan R.R., and Tufarelli C. Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer. Nucleic Acids Res. 2013; 41:6857–6869

- Das S., Lohia A. Delinking of S phase and cytokinesis in the protozoan parasite *Entamoeba histolytica*. Cell Microbiol. 2002; 4(1):55-60.

- Dastidar P.G., Majumder S., Lohia A. Eh Klp5 is a divergent member of the kinesin 5 family that regulates genome content and microtubular assembly in *Entamoeba histolytica*. Cell Microbiol. 2007; 9(2):316-28.

- Davis C.M., Constantinides P.G., van der Riet F *et al*. Activation and demethylation of the intracisternal A particle genes by 5-azacytidine. Cell Differ Dev. 1989; 27:83–93.

- Davis P.H., Zhang Z., Chen M., Zhang X., Chakraborty S., Stanley S.L. Jr. Identification of a family of BspA like surface proteins of *Entamoeba histolytica* with novel leucine rich repeats. Mol Biochem Parasitol. 2006; 145(1):111-6.

- Dawson A., Hartswood E., Paterson T., Finnegan D.J.: A LINE-like transposable element in *Drosophila*, the *I* factor, encodes a protein with properties similar to those of retroviral nucleocapsids**.** EMBO J. 1997; 16:4448-4455.

- De la Vega H, Specht C.A, Chau A, Semino C.E., Robbins P.W., Eichinger D., Caplivski D., Ghosh S., Samuelson J. Cyst-specific exochitinases of *Entamoebae* contain unique hydrophilic repeats at their amino termini. Arch Med Res. 1997; 28:143-6.

- DeBerardinis R.J., Kazazian H.H. Jr. Analysis of the promoter from an expanding mouse retrotransposon subfamily**.** *Genomics.*1999; 56:317-323.

- Deininger P., Belancio V.P. Detection of LINE-1 RNAs by Northern Blot. Methods Mol Biol. 2016; 1400:223-36.

- Deininger P.L., Moran J.V., Batzer M.A., and Kazazian H.H. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 2003; 13:651-658.

- Charcas-Lopez Mdel S., Garcia-Morales L., Pezet-Valdez M., Lopez-Camarillo C., Zamorano-Carrillo A., Marchat L.A. Expression of EhRAD54, EhRAD51, and EhBLM proteins during DNA repair by homologous recombination in *Entamoeba histolytica*. Parasite. 2014; 21:7.

- Demple B., and Harrison L. Repair of oxidative damage to DNA: enzymology and biology. Annu Rev Biochem. 1994; 63:915-948.

- Derek Huntley M., Ioannis Pandis, Sarah Butcher A. and John Ackers P. Research article Bioinformatic analysis of *Entamoeba histolytica* SINE1 elements. BMC Genomics. 2010; 11:321.

- Dewannieux M., and Heidmann T. Role of poly(A) tail length in Alu retrotransposition. Genomics. 2005; 86;378-381.

- Dewannieux, M., Esnault C., and Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 2003; 35:41-48.

- Dhar S.K., Choudhury N.R., Bhattacharya A., Bhattacharya S. A multitude of circular DNAs exist in the nucleus of *Entamoeba histolytica*. Mol Biochem Parasitol. 1995; 70(1-2):203-6.

- Dombroski B.A., Feng Q., Mathias S.L., Sassaman D.M., Scott A. F., Kazazian H.H., Jr and Boeke J.D. Mol. Cell. Biol. 1994; 14:4485–4492.

- Doucet A.J., Hulme A.E., Sahinovic E., Kulpa D.A., Moldovan J.B., Kopera H.C., Athanikar J.N., Hasnaoui M., Bucheton A., Moran J.V., et al. Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet 2010; 6; PMID:20949108;

- Doucet A.J., Wilusz J.E., Miyoshi T., Liu Y., Moran J.V. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. Mol Cell. 2015; 60(5):728-41.

- Du T. & Zamore P.D. MicroPrimer: the biogenesis and function of microRNA. Development. 2005; 132:4645-4652.

- Dudley J.P. Discrete high molecular weight RNA transcribed from the long interspersed repetitive element L1Md. Nucleic Acids Res. 1987; 15(6):2581-92.

- Duggal P., Guo X., Haque R., *et al*. A mutation in the leptin receptor is associated with *Entamoeba histolytica* infection in children. J Clin Invest. 2011; 121(3):1191-8.

- Gogvadze E. and Buzdin A. Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci. 2009; 66:3727-3742.

- Eichinger L., Pachebat J.A., Glöckner G., *et al*. The genome of the social amoeba *Dictyostelium discoideum*. Nature. 2005; 435(7038):43-57.

- Eickbush T.H., Malik H.S. Origins and evolution of retrotransposons**.** In *Mobile DNA II* Edited by: Craig NL, Craigie R, Gellert M, Lambowitz AM. Washington, DC: ASM Press; 2002:1111-1144

- Eickbush T.H. R2 and related site-specific non-long terminal repeat retrotransposons, In N. L. Craig, R. Craigie, M. Gellert, and A.M. Lambowitz (ed.), Mobile DNA II. American Society for Microbiology. 2002; 813-835.

- Eickbush T.H., and Jamburuthugoda V.K. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 2008; 134: 221-234.

- Ergün S., Buschmann C., Heukeshoven J., Dammann K., Schnieders F., Lauke H., Chalajour F., Kilic N., Strätling W.H., Schumann G.G. Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. J Biol Chem 2004; 279:27753-63;

- Esnault C., Maestre J., and Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000; 24:363-367.

- Evgen'ev M.B., Arkhipova I.R. Penelope-like elements—a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res. 2005; 110:510-521.

- Evgen'ev M.B., Zelentsova, H., Shostak, N., Kozitsina, M., Barskyi, V., Lankenau, D.H., and Corces, V.G. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. Proc Natl Acad Sci U S A. 1997; 94:196-201.

- Faghihi M.A., Wahlestedt C. Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol. 2009; 10(9):637-43.

- Faghihi M.A. *et al.* Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol.2010; 11: R56.

- Faulkner G.J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009 May;41(5):563-71.

- Fawcett D.H., Lister C.K., Kellett E., and Finnegan D.J. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs. Cell. 1986; 47:1007-1015.

- Feng Q., Moran J.V., Kazazian H.H. Jr, Boeke J.D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition**.** Cell 1996, 87:905-916.

- Feng Q., Schumann G., and Boeke J.D. Retrotransposon R1Bm endonuclease cleaves the target sequence. Proc Natl Acad Sci U S A. 1998; 95:2083-2088.

- Feschotte C., Pritham E.J. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007; 41:331-368.

- Feschotte C, Pritham E.J. Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet. 2005; 21:551–552

- Feschotte C. Transposable elements and the evolution of regulatory networks, Nat. Rev. Genet. 2008; 9: 397–405.

- Filichkin S.A., Priest H.D., Givan S.A., Shen R., Bryant D.W., Fox S.E., Wong W.K. and Mockler T.C. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. 2010; 20:45-58.

- Filipowicz W., Jaskiewicz L., Kolb F.A. & Pillai R.S. Post-transcriptional gene silencing by siRNAs and miRNAs. Curr. Opin. Struct. Biol. 2005; 15:331–341.

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

- Garcia-Perez J.L., Gonzalez C.I., Thomas M.C., Olivares M., and Lopez M.C. Characterization of reverse transcriptase activity of the L1Tc retroelement from *Trypanosoma cruzi*. Cell Mol Life Sci. 2003; 60:2692-2701.

- Gardner M.J., Shallom S.J., Carlton J.M., *et al*. Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature. 2002; 419(6906):531-4.

- Gasior S.L., Wakeman T.P., Xu B., Deininger P.L. The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol. 2006; 357:1383–1393.

- Gaurav A.K., Kumar J., Agrahari M., Bhattacharya A., Yadav V.P., Bhattacharya S. 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. 2016; 211:84-93.

- Gelderman A.H., Bartgis I.L., Keister D.B., Diamond L.S. A comparison of genome sizes and thermal-denaturation-derived base composition of DNAs from several members of *Entamoeba* (*histolytica* group). J Parasitol. 1971; 57(4):912-6.

- Gelderman A.H., Keister D.B., Bartgis I.L., Diamond L.S. Characterization of the deoxyribonucleic acid of representative strains of *Entamoeba histolytica*, *E. histoytica* like amebae, and *E. moshkovskii*. J Parasitol. 1971; 57(4):906-11.

- Gentles A.J., Wakefield M.J., Kohany O., Gu, W., Batzer M.A., Pollock D.D., and Jurka J. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. Genome Res. 2007; 17**:**992-1004.

- George J.A., Eickbush T.H. Conserved features at the 5 end of *Drosophila* R2 retrotransposable elements: implications for transcription and translation**.** Insect Mol Biol. 1999, 8:3-10.

- Ghosh S., Field J., Rogers R., Hickman M., Samuelson J. The *Entamoeba histolytica* mitochondrion-derived organelle (crypton) contains double-stranded DNA and appears to be bound by a double membrane. Infect Immun. 2000; 68(7):4319-22.

- Girard-Misguich F., Sachse M., Santi-Rocca J., Guillén N. The endoplasmic reticulum chaperone calreticulin is recruited to the uropod during capping of surface receptors in *Entamoeba histolytica*. Mol Biochem Parasitol. 2008; 157(2):236-40.

- Gogvadze E., and Buzdin, A. Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci. 2009; 66,3727-3742.

- Gonzalez-Salazar F., Mata-Cárdenas B.D., Vargas-Villareal J. [Sensibility of *Entamoeba histolytica* trophozoites to ivermectin]. Medicina (B Aires). 2009; 69(3):318-20.

- Goodier J.L., Ostertag E.M., Engleka K.A., Seleme M.C., Kazazian H.H. Jr. A potential role for the nucleolus in L1 retrotransposition. Hum Mol Genet 2004; 13:1041-8.

- Goodier J.L., Zhang L., Vetter M.R. and Kazazian H. H. Jr. LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. Mol. Cell. Biol. 2007; 27, 6469–6483.

- Goodwin T.J., Poulter R.T. A new group of tyrosine recombinase encoding retrotransposons. Mol Biol Evol. 2004; 21:746–759.

- Gregory J. Cost, Qinghua Feng, Alain Jacquier and Jef D. Boeke. Human L1 element target-primed reverse transcription in vitro. The EMBO Journal. 2002; 21(21):5899–5910.

- Grewal J.S., Padhan N., Aslam S., Bhattacharya A, Lohia A. The calcium binding protein EhCaBP6 is a microtubular-end binding protein in *Entamoeba histolytica*. Cell Microbiol. 2013; 15(12):2020-33.

- Grimaldi G., Skowronski J., and Singer M.F. Defining the beginning and end of KpnI family segments. EMBO J. 1984; 3:1753-1759.

- Guibert S., Forné T., weber M.. Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome Res. 2012; 22:633-641.

- Haas N.B., Grabowski J.M., North J., Moran J.V., Kazazian H.H., Burch J.B. Subfamilies of CR1 non-LTR retrotransposons have different 5'UTR sequences but are otherwise conserved**.** Gene.2001; 265:175-183.

- Hackett J.A., Reddington J.P., Nestor Ce et al. Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline. Development. 2012; 139:3623-3632.

- Hamano S., Becker S., Asgharpour A., Ocasio Y.P., Stroup S.E., McDuffie M., Houpt E. Gender and genetic control of resistance to intestinal amebiasis in inbred mice. Genes Immun. 2008; 9(5):452-61.

- Han J.S., Szak S.T., Boeke J.D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004;429(6989):268-74.

- Hancks D.C., Kazazian H.H. Active human retrotransposons: Variation and disease. Curr Opin Genet Dev. 2012; 22(3):191-203.

- Haque R., Huston C.D., Hughes M., Houpt E., and Petri W.A., Jr. Amebiasis. N Engl J Med 2003; 348:1565-1573.

- Harris D., Yadav P.N., Pandey V.N. Loss of polymerase activity due to Tyr to Phe substitution in the YMDD motif of human immunodeficiency virus type-1 reverse transcriptase is compensated by Met to Val substitution within the same motif. Biochemistry. 1998; 37:9630-9640.

- Hata K., Sakaki Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene. 1997; 189(2):227–34.

- Heras S. R., Lopez M.C., Olivares M., Thomas M.C. 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. 2007; 35(7):2199-214.

- Hohjoh H., Singer M.F. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon**.** EMBO J. 1997; 16:6034-6043.

- Hohjoh H., and Singer M.F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. Embo J. 1996; 15:630-639.

- Holmes S.E., Dombroski B.A., Krebs C.M., Boehm C.D. and Kazazian H.H., Jr. Nature Genet. 1994; **7:**143–148

- Hon C.C. *et al*. Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. Nucleic Acids Res. 2013; 41(3):1936-52.

- Hu J., Qin H., Sharma M., Cross T.A., Gao F.P. Chemical cleavage of fusion proteins for high-level production of transmembrane peptides and protein domains containing conserved methionines. Biochim Biophys Acta. 2008; 1778(4):1060-6.

- Huston C.D. Parasite and host contributions to the pathogenesis of amebic colitis. Trends Parasitol. 2004; 20:23-26.

- Hackett J.A. Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline. Development. 2012; 139:3623–3632.

- Law J.A. Establishing, maintaining and modifying DNA methylation patterns in plants and animals, Nat. Rev. Genet. 2010; 11:204–220.

- Yoder J.A. Cytosine methylation and the ecology of intragenomic parasites, Trends Genet. 1997; 13:335–340.

- Crichton J.H. Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline, Cell. Mol. Life Sci. 2014; 71 (9): 1581–1605.

- Jackson-Grusby L., Beard C., Possemato R. *et al*. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nat Genet. 2001; 27:31-39.

- Jakubczak J.L., Burke W.D., and Eickbush T.H. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proc Natl Acad Sci U S A. 1991; 88:3295-3299.

- Jhingan G.D., Panigrahi S.K., Bhattacharya A., Bhattacharya S. The nucleolus in *Entamoeba histolytica* and *Entamoeba invadens* is located at the nuclear periphery. Mol Biochem Parasitol. 2009; 167(1):72-80.

- Jorge Cruz-Reyes, Tayyab ur-Rehman, William Spice M., John Ackers P. A novel transcribed repeat element from *Entamoeba histolytica.* Gene. 1995; 166:183-184.

- Kajikawa M., and Okada N. LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell. 2002; 111:433-444.

- Kaneda M., Okano M., Hata, K., Sado T., Tsujimoto N., Li E., and Sasaki H. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature. 2004; 429, 900–903.

- Kapitonov V.V., Jurka J. Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci USA. 2006; 103:4540-4545.

- Kapitonov V.V., Tempel S., Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene. 2009; 448(2):207-13.

- Kaplan W., Hüsler P., Klump H., Erhardt J., Sluis-Cremer N., Dirr H. Conformational stability of pGEX-expressed Schistosoma japonicum glutathione S-transferase: a detoxification enzyme and fusion-protein affinity tag. Protein Sci. 1997 Feb;6(2):399-406.

- Katayama S. *et al*. Antisense transcription in the mammalian transcriptome. Science. 2005; 309:1564–1566.

- Kaushik N., Chowdhury K., Pandey V.N., and Modak M.J. Valine of the YVDD motif of moloney murine leukemia virus reverse transcriptase: role in the fidelity of DNA synthesis. Biochemistry. 2000; 39:5155-5165.

- Kawano M., Aravind L. & Storz G. An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. Mol. Microbiol. 2007; 64:738–754.

- Keren H., Lev-Maor G. and Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. 2010; 11: 345-355.

- Khan H., Smit A., Boissinot S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates**.** Genome Res. 2006; 16:78-87.

- Khazina E. and Weichenrieder O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A. 2009; 106: 731-736.

- Kidwell M.G., Lisch D. Transposable elements as sources of genomic variation. Moblie DNAII. Washington (DC): American Society for Microbiology Press. 2002; p. 59–90.

- Kiedzierska A., Czepczynska H., Smietana K., Otlewski J. Expression, purification and crystallization of cysteine-rich human protein muskelin in *Escherichia coli*. Protein Expr Purif. 2008; 60(1):82-8.

- Kirilyuk A., Tolstonog G.V., Damert A., Held U., Hahn S., Löwer R., Buschmann C., Horn A. V., Traub P., Schumann G.G. Functional endogenous LINE-1 retrotransposons are expressed and mobilized in rat chloroleukemia cells. Nucleic Acids Res. 2008; 36:648-65.

- Kohlstaedt L.A., Wang J., Friedman J.M., Rice P.A., and Steitz T.A. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science. 1992; 256:1783-1790.

- Kojima K.K., and Fujiwara H. An extraordinary retrotransposon family encoding dual endonucleases. Genome Res. 2005; 15:1106-1117.

- Kolosha V.O., Martin S.L. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1)**.** *J Biol Chem.* 2003; 278:8112-8117.

- Kolosha V.O. and Martin S.L. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A. 1997; 94:10155-10160.

- Kramerov D.A., Vassetzky N.S. Short retroposons in eukaryotic genomes**.** Int Rev Cytol 2005, 247:165-221.

- Kulpa D.A., and Moran J.V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol. 2006; 13:655-660.

- Kurzynska-Kokorniak A., Jamburuthugoda V.K., Bibillo, A., and Eickbush T.H. DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. J Mol Biol. 2007; 374:322-333.

- Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. *et al*. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860-921.

- Lane N., Dean W., Erhardt S. *et al*. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis. 2003; 35:88-93.
- Larder BA, Purifoy DJ, Powell KL, Darby G: Site-specific mutagenesis of AIDS virus reverse transcriptase. Nature. 1987; 327:716-717.

- Lasa I. *et al.* Genome-wide antisense transcription drives mRNA processing in bacteria. Proc. Natl Acad. Sci. USA. 2011; 108:20172-20177.

- Leng N. and Kendziorski C. EBSeq: An R package for gene and isoform differential expression analysis of RNA-Seq data. 2015;

- León-Avila G., Tovar J. Mitosomes of *Entamoeba histolytica* are abundant mitochondrion-related remnant organelles that lack a detectable organellar genome. Microbiology. 2004; 150:1245-50.

- Li B., Dewey C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323.

- Li E., Bestor T.H. and Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992; 69:915-926.

- Lingner J., Hughes T.R., Shevchenko A., Mann M., Lundblad V. and Cech T.R. Reverse transcriptase motifs in the catalytic subunit of telomerase. Science. 1997; 276:561–567.

- Lioutas C., Tannich E. Transcription of protein-coding genes in *Entamoeba histolytica* is insensitive to high concentrations of alpha-amanitin. Mol Biochem Parasitol. 1995; 73(1-2):259-61.

- Lister R., Pelizzola M., Dowen R.H. *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315-322.

- Dai L., LaCava J., Taylor M.S. and Boeke J.D. Expression and detection of LINE-1 ORF-encoded proteins. mobile Genetic elements. 2014, 22(4): 29319.

- Lobstein J., Emrich C.A., Jeans C., Faulkner M., Riggs P., Berkmen M. SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. Microb Cell Fact. 2016; 5(1):124.

- Loftus B., Anderson I., Davies R. *et al.* The genome of the protist parasite *Entamoeba histolytica*. Nature. 2005; 433(7028):865-8.

- Lohia A., Haider N., Biswas B. B. Characterisation of a repetitive DNA family from *Entamoeba histolytica* containing *Saccharomyces cerevisiae* ARS consensus sequences. Gene. 1990; 96(2):197-203.

- Lohia A., Mukherjee C., Majumder S., Dastidar P.G. Genome re-duplication and irregular segregation occur during the cell cycle of *Entamoeba histolytica*. Biosci Rep. 2007; 27(6):373-84.

- López-Casamichana M., Orozco E., Marchat L.A., López-Camarillo C. Transcriptional profile of the homologous recombination machinery and characterization of the EhRAD51 recombinase in response to DNA damage in *Entamoeba histolytica*. BMC Mol Biol. 2008; 9:35.

- Lorenzi H., Thiagarajan M., Haas B., Wortman J., Hall N., Caler E. Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. BMC Genomics. 2008; 9:595.

- Lorenzi H.A., Puiu D., Miller J.R., Brinkac L.M., Amedeo P., Hall N., Caler E.V. New assembly, reannotation and analysis of the *Entamoeba histolytica* genome reveal new genomic features and protein content information. PLoS Negl Trop Dis. 2010; 4(6): e716.

- Losch F. Massenhafte Entwickelung Von Amöben Im Dickdarm. Arch F Path Anat 1875; 65: 196–211.

- Luan, D.D., and Eickbush, T.H. Downstream 28S gene sequences on the RNA template affect the choice of primer and the accuracy of initiation by the R2 reverse transcriptase. Mol Cell Biol. 1996; 16:4726-4734.

- Luan, D.D., and Eickbush, T.H. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol Cell Biol. 1995; 15:3882-3891.

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

- Ludvík J., Shipstone A.C. The ultrastructure of *Entamoeba histolytica*. Bull World Health Organ. 1970; 43(2):301-8.

- Fraga M.F., Esteller M. DNA methylation: a profile of methods and applications, BioTechniques 33 (2002) 632–649.

- Macias F., Lopez M.C., Thomas M.C. The Trypanosomatid Pr77 hallmark contains a downstream core promoter element essential for transcription activity of the *Trypanosoma cruzi* L1Tc retrotransposon. BMC Genomics. 2016; 17:105.

- Mackey-Lawrence N.M., Guo X., Sturdevant D.E. *et al*. Effect of the leptin receptor Q223R polymorphism on the host transcriptome following infection with *Entamoeba histolytica*. Infect Immun. 2013; 81(5):1460-70.

- MacLeod A., Tweedie A., McLellan S. *et al*. The genetic map and comparative analysis with the physical map of *Trypanosoma brucei*. Nucleic Acids Res. 2005; 33(21):6688-93.

- Majumder S., Lohia A. *Entamoeba histolytica* encodes unique formins, a subset of which regulates DNA content and cell division. Infect Immun. 2008; 76(6):2368-78.

- Maksakova I.A., Romanish M.T., Gagnier L., Dunn C.A., Van de Lagemaat L.N., and Mager D.L. Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *PLoS Genet.* 2006;

- Malhotra A. Tagging for protein expression. Methods Enzymol. 2009; 463:239-58.

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

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

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

- Mandal P.K., Rawal K., Ramaswamy R., Bhattacharya A., Bhattacharya S. Identification of insertion hot spots for non-LTR retrotransposons: computational and biochemical application to *Entamoeba histolytica*. Nucleic Acids Res. 2006; 34(20):5752-63.

- Mandal P.K., Bagchi A., Bhattacharya A., and Bhattacharya S. An *Entamoeba histolytica* LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease. Eukaryot Cell. 2004; 3:170-179.

- Martín Caballero I., Hansen J., Leaford D. *et al.* The methyl-CpG binding proteins Mecp2, Mbd2 and Kaiso are dispensable for mouse embryogenesis, but play a redundant function in neural differentiation. PLoS One. 2009; 4: e4315.

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

- Martin F., Maranon C., Olivares M., Alonso C., and Lopez M.C. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. J Mol Biol. 1995; 247:49-59.

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

- Martin S.L. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol. 1991; 11:4804-4807.

- Mathias S.L., Scott A.F., Kazazian H.H., Jr., Boeke J.D., and Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991; 254:1808-1810.

- Matlik K., Redik K., and Speek M. L1 antisense promoter drives tissue-specific transcription of human genes. J. Biomed. Biotechnol. 2006; 71753.

- McLean C., Bucheton A., Finnegan D.J. The 5' untranslated region of the *I* factor, a long interspersed nuclear element-like retrotransposon of *Drosophila melanogaster*, contains an internal promoter and sequences that regulate expression**.** *Mol Cell Biol.* 1993; 13:1042-1050.

- Medstrand P., Van de Lagemaat L.N., Mager D. L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 2002; 12:1483-1495.

- Meissner A., Mikkelsen T.S., Gu H. *et al*. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454:766–770.

- Michel B., Alagón A., Lizardi P.M., Zurita M. Characterization of a repetitive DNA element from *Entamoeba histolytica*. Mol Biochem Parasitol. 1992; 51(1):165-8.

- Mikkelsen T.S., Wakefield M.J., Aken B., Amemiya C.T., Chang J.L., Duke S., Garber M., Gentles A.J., Goodstadt L., Heger A. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature*. 2007; 447**:**167-177.

- Miller J.H., Swartzwelder J.C., Deas J.E. An electron microscopic study of *Entamoeba histolytica*. J Parasitol. 1961; 47:577-87.

- Millican D.S., Bird I.M. A general method for single-stranded DNA probe generation. Anal Biochem. 1997 Jun 15; 249(1):114-7.

- Minakami R., Kurose K., Etoh K., Furuhata Y., Hattori M., Sakaki Y. Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element**.** *Nucleic Acids Res.* 1992; 20:3139-3145.

- Minchiotti G., Di Nocera P. P. Convergent transcription initiates from oppositely oriented promoters within the 5' end regions of *Drosophila melanogaster* F elements**.** Mol Cell Biol. 1991; 11:5171-5180.

- Mirelman D., Anbar M., Bracha R. Epigenetic transcriptional gene silencing in *Entamoeba histolytica***.** IUBMB Life. 2008; 60:598-604.

- Mittal V., Bhattacharya A., Bhattacharya S. Isolation and characterization of a species-specific multicopy DNA sequence from *Entamoeba histolytica*. Parasitology. 1994; 108:237-44.

- Mittal V., Ramachandran S., Sehgal D., Bhattacharya A., Bhattacharya S. Sequence analysis of a DNA fragment with yeast autonomously replicating sequence activity from the extrachromosomal ribosomal DNA circle of *Entamoeba histolytica*. Nucleic Acids Res. 1991; 19(10):2777.

- Mittal V., Sehgal D., Bhattacharya A., Bhattacharya S. A second short repeat sequence detected downstream of rRNA genes in the *Entamoeba histolytica* rDNA episome. Mol Biochem Parasitol. 1992; 54(1):97-100.

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

- Moran J.V., Holmes S.E., Naas T.P., DeBerardinis R.J., Boeke J.D. and Kazazian H.H., Jr. High-frequency retrotransposition in cultured mammalian cells. Cell. 1996; 87:917-927.

- Moran J.V., DeBerardinis R.J. & Kazazian H.H. Jr. Exon shuffling by L1 retrotransposition. Science.1999; 283:1530–1534.

- Morrish T.A., Garcia-Perez J.L., Stamato T.D., Taccioli G.E, Sekiguchi J., Moran J.V. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature. 2007; 446(7132):208-12.

- Mount S.M. and Rubin G.M. Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol Cell Biol. 1985; 5:1630-1638.

- Mourier T, Willerslev E. Does selection against transcriptional interference shape retroelement-free regions in mammalian genomes? PLoS ONE. 2008; 3: e3760.

- Mukherjee C., Majumder S., Lohia A. Inter-cellular variation in DNA content of *Entamoeba histolytica* originates from temporal and spatial uncoupling of cytokinesis from the nuclear cycle. PLoS Negl Trop Dis. 2009; 3(4): e409.

- Muotri A.R., Marchetto M.C.N., Coufal N.G. *et al*. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010; 468:443-446.

- Nakamura T.M., Morin G.B., Chapman K.B., Weinrich S.L., Andrews W.H., Lingner J., Harley C.B. and Cech T.R. Telomerase catalytic subunit homologs from fission yeast and human. Science. 1997; 277: 955–959.

- Nallamsetty S., Waugh D. S. Mutations that alter the equilibrium between open and closed conformations of *Escherichia coli* maltose-binding protein impede its ability to enhance the solubility of passenger proteins. Biochem Biophys Res Commun. 2007 Dec 21;364(3):639-44.

- Nan X., Ng H.H., Johnson C.A. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature. 1998; 393:386-389.

- Nigumann P., Redik K., Matlik K., Speek M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics. 2002; 79:628-634.

- Fisher O., Pleiotropic phenotype in *Entamoeba histolytica* overexpressing DNA methyltransferase (Ehmeth), Mol. Biochem. Parasitol. 2006; 147:48–54.

- Ohshima K. and Okada N. Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. J Mol Biol. 1994; 243:25-37.

- Ohshima K., Hamada M., Terai Y. and Okada N. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Mol Cell Biol. 1996; 16:3756-3764.

- Ohshima K., Hattori M., Yada T., Gojobori T., Sakaki Y., and Okada N. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 2003; 4: R74.

- Ohshima K., Hamada M., Terai Y. and Okada N. The 3 ends of tRNA-derived short interspersed repetitive elements are derived from the 3 ends of long interspersed repetitive elements. Mol. Cell. Biol. 1996; 16**:**3756-3764.

- Okada N. and Hamada M. The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. J Mol Evol. 1997; 44 Suppl 1, S52-56.

- Okano M., Bell D.W., Haber D.A., Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999; 99:247–257.

- Okazaki S., Ishikawa H., and Fujiwara H. Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. Mol Cell Biol. 1995; 15:4545-4552.

- Olivares M., Garcia-Perez J.L., Thomas M.C., Heras S.R. and Lopez M.C. The non-LTR (long terminal repeat) retrotransposon L1Tc from *Trypanosoma cruzi* codes for a protein with RNase H activity. J Biol Chem. 2002; 277:28025-28030.

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

- Öllinger R., Reichmann J., Adams I.R. Meiosis and retrotransposon silencing during germ cell development in mice. Differentiation. 2010; 79:147-158.

- Orozco E., Solís F.J., Domínguez J., Chávez B., Hernández F. *Entamoeba histolytica*: cell cycle and nuclear division. Exp Parasitol. 1988; 67(1):85-95.

- Ostertag E.M., Goodier J.L., Zhang Y. and Kazazian Jr., H.H. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum.* Genet. 2003; 73:1444-1451.

- Ozsolak F., Kapranov P., Foissac S., Kim S.W., Fishilevich E., Monaghan A.P., John B. and Milos P.M. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell. 2010; 143:1018-1029.

- Pan Q., Shai O., Lee L.J., Frey B.J. and Blencowe B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 2008; 40: 1413–1415.

- Pauws E., Van Kampen A.H., Van de Graaf S.A., De Vijlder J.J. and Ris-Stalpers C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res. 2001; 29:1690-1694.

- Perepelitsa-Belancio V., Deininger P. RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat Genet. 2003; 35:363-366.

- Pickeral O.K., Makalowski W., Boguski M.S., Boeke J.D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 2000; 10:411-415.

- Piskareva O., Schmatchenko V. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. FEBS Lett. 2006 Jan 23; 580(2):661-8.

- Poch O., Sauvaget I., Delarue M., and Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. Embo J. 1989; 8:3867-3874.

- Popp C., Dean w., Feng S. *et al*. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature. 2010; 463:1101-1105.

- Poulter R.T. and Goodwin T.J. DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res. 2005; 110:575-588.

- Pritham E.J., Feschotte C., Wessler S.R. Unexpected diversity and differential success of DNA transposons in four species of *Entamoeba* protozoans. Mol Biol Evol. 2005; 22(9):1751-63.

- Pritham E.J., Putliwala T., Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007; 390:3-17.

- R Pelizzola M., Dowen R.H. *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322.

- Ollinger R. Meiosis and retrotransposon silencing during germ cell development in mice, Differentiation. 2010; 79 (3):147–158.

- Elbarbary R.A. *et al*. Retrotransposons as regulators of gene expression, Science. 2016; 351 (6274) aac7247, 12.

- Reddington J.P., Pennings S., Meehan R.R. Non-canonical functions of the DNA methylome in gene regulation. Biochem J. 2013; 451:13-23.

- Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007; 447:425–432.

- Richard Cordaux and Mark Batzer A. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009 Oct; 10(10): 691–703.

- Richardson S.R., Doucet A.J., Kopera H.C., Moldovan J.B., Garcia-Perez J. L. and Moran J. V. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol. 2015; Spectr. 3, MDNA3-0061-2014.

- Rollins R.A., Haghighi F., Edwards J.R., Das R Zhang M. Q., Ju J. *et al*. Largescale structure of genomic methylation patterns. Genome Res. 2006; 16(2):157–63.

- Romanish M.T., Cohen C.J., Mager D.L. Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer. Semin Cancer Biol. 2010; 20:246–253.

- Rosano G.L., Ceccarelli E.A. Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. 2014; 17(5)172.

- Rowe H.M., Trono D. Dynamic control of endogenous retroviruses during development. Virology. 2011; 411:273-287.

- Saab B.R., Musharrafieh U., Nassar N.T., Khogali M., Araj G.F. Intestinal parasites among presumably healthy individuals in Lebanon. Saudi Med J. 2004; 25(1):34-7

- Salit I.E., Khairnar K., Gough K., Pillai D.R. A possible cluster of sexually transmitted *Entamoeba histolytica*: genetic analysis of a highly virulent strain. Clin Infect Dis. 2009; 49(3):346-53.

- SanMiguel P., Gaut B. S., Tikhonov A., Nakajima Y. and Bennetzen J. L. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998; 20:43-45.

- Sargeaunt P.G. and Williams J.E. Electrophoretic isoenzyme patterns of *Entamoeba histolytica* and *Entamoeba coli*. Trans R Soc Trop Med Hyg. 1978; 72:164-166.

- Sassaman D.M., Dombroski B.A., Moran J.V., Kimberland M.L., Naas T.P., DeBerardinis R.J., Gabriel A., Swergold G.D. and Kazazian H.H., Jr Nature Genet. 1997; 16:37–43

- Satish S., Bakre A.A., Bhattacharya S., Bhattacharya A. Stress-dependent expression of a polymorphic, charged antigen in the protozoan parasite *Entamoeba histolytica*. Infect Immun. 2003; 71(8):4472-86.

- Schaudinn F. Untersuchungen uber die fortpflanzung einiger rhizopoden. Arbeiten aus dem kaiserlichen gesundheitsamte. 1903; 19:547-576.

- Schostak N., Pyatkov K., Zelentsova E., Arkhipova I., Shagin D., Shagina I., Mudrik E., Blintsov A., Clark I., Finnegan D.J., and Evgen'ev M. Molecular dissection of Penelope transposable element regulatory machinery. Nucleic Acids Res. 2008; 36:2522-2529.

- Schumann G., Zundorf I., Hofmann J., Marschalek R., Dingermann T. Internally located and oppositely oriented polymerase II promoters direct convergent transcription of a LINE-like retroelement, the *Dictyostelium* repetitive element, from *Dictyostelium discoideum***.** Mol Cell Biol.1994; 14:3074-3084.

- Seisenberger S., Andrews S., Krueger F. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. Mol Cell. 2012; 48:849-862.

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

- Sigova A.A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl Acad. Sci. USA. 2013; 110:2876-2881.

- Singh N., Bhattacharya A., Bhattacharya S. Homologous recombination occurs in *Entamoeba* and is enhanced during growth stress and stage conversion. PLoS One.2013; 8(9): e74465.

- Slotkin R.K., Freeling M. and Lisch D. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat Genet. 2005; 37:641-644.

- Smit A.F. and Riggs A.D. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. 1995; 23:98-102.

- Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol. 2001; 21:1973-1985.

- Stanley S.L Jr. Amoebiasis. Lancet. 2003; 361(9362):1025-34.

- Stanley S.L. Pathophysiology of amoebiasis. Trends Parasitol. 2001; 17(6):280-5

- Sverdlov E.D. Retroviruses and primate evolution. Bioessays. 2000; 22:161-171.

- Swergold G.D. Identification, characterization, and cell specificity of a human LINE-1 promoter**.** Mol Cell Biol. 1990; 10:6718-6729.

- Szak S.T., Pickeral O.K., Makalowski W., Boguski M.S., Landsman D., Boeke J.D. Molecular archeology of L1 insertions in the human genome**.** Genome Biol. 2002; 3: research0052.

- Lavi T. Sensing DNA methylation in the protozoan parasite *Entamoeba histolytica*, Mol. Microbiol. 2006; 62:1373–1386.

- Tachibana H., Yanagi T., Akatsuka A., Kobayashi S., Kanbara H., Tsutsumi V. Isolation and characterization of a potentially virulent species *Entamoeba nuttalli* from captive Japanese macaques. Parasitology. 2009; 136(10):1169-77.

- Tachibana H., Yanagi T., Pandey K., Cheng X.J., Kobayashi S., Sherchand J.B., Kanbara H. An *Entamoeba* sp. strain isolated from rhesus monkey is virulent but genetically different from *Entamoeba histolytica*. Mol Biochem Parasitol. 2007; 153(2):107-14.

- Takahashi H., Okazaki S. and Fujiwara H. A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori. Nucleic Acids Res. 1997; 25:1578-1584.

- Tang W., Luo X.Y., Sanmuels V. Gene silencing: double-stranded RNA mediated mRNA degradation and gene inactivation. Cell Res. 2001;11(3):181-6.

- Thayer R.E., Singer M.F., Fanning T.G. Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1 -encoded protein. Gene. 1993;133(2):273–7.

- Tovar J., Fischer A., Clark C.G. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica*. Mol Microbiol. 1999; 32(5):1013- 21.

- Trelogan S.A. and Martin S.L. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc Natl Acad Sci U S A. 1995; 92:1520-1524.

- Ullu E. and Tschudi C. Alu sequences are processed 7SL RNA genes. Nature. 1984; 312:171-172.

- Yadav V.P. *et al*. Recombinant SINEs are formed at high frequency during induced retrotransposition in vivo, Nat. Commun. 2012; 22(3):854.

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

- Vanin E.F. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 1985; 19: 253-272.

- Walsh J.A. Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality. Rev Infect Dis. 1986; 8(2):228-38

- Walsh C.P. and Bestor T.H. Cytosine methylation and mammalian development. *Genes Dev*. 1999; 13:26-34.

- Wang E.T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., Kingsmore S.F., Schroth G.P. and Burge C.B. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470-476.

- Watanabe T. *et al*. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008; 453(7194):539-43.

- Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M. An P *et al.* Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520-562.

- Weber C., Blazquez S., Marion S., Ausseur C., Vats D., Krzeminski M., Rigothier M.C., Maroun R.C., Bhattacharya A., Guillén N. Bioinformatics and functional analysis of an *Entamoeba histolytica* mannosyltransferase necessary for parasite complement resistance and hepatical infection. PLoS Negl Trop Dis. 2008; 2(2): e165.

- Wei W., Gilbert N., Ooi S.L., Lawler J.F., Ostertag E.M., Kazazian H.H., Boeke J.D., Moran J.V. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 2001; 21(4):1429-39.

- Weiner A.M., Deininger P.L. and Efstratiadis A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem. 1986; 55:631-661.

- Weinke T., Friedrich-Jänicke B, Hopp P, Janitschke K. Prevalence and clinical importance of *Entamoeba histolytica* in two high-risk groups: travelers returning from the tropics and male homosexuals. J Infect Dis. 1990; 161(5):1029-31.

- Wessler S.R. Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci USA. 2006; 103:17600–17601.

- Wheelan S.J., Aizawa Y., Han J.S., Boeke J.D. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res. 2005; 15:1073-1078.

- Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; 8:973-982.

- Wicker T. and Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. 2007; 17: 1072-1081.

- Wickstead B., Ersfeld K., Gull K. Repetitive elements in genomes of parasitic protozoa. Microbiol Mol Biol Rev. 2003; 67(3):360-75

- Wiench M., John S., Baek S. *et al.* DNA methylation status predicts cell type-specifc enhancer activity. eMBO J. 2011; 30:3028-3039.

- Wilhoeft U., Bub H. and Tannich E. The abundant polyadenylated transcript 2 DNA sequence of the pathogenic protozoan parasite *Entamoeba histolytica* represents a nonautonomous non-long-terminal-repeat retrotransposon-like element which is absent in the closely related nonpathogenic species *Entamoeba dispar*. Infect. Immun. 2002; 70**:**6798-6804.

- Wilhoeft U., Bub H. and Tannich E. Analysis of cDNA expressed sequence tags from *Entamoeba histolytica*: identification of two highly abundant polyadenylated transcripts with no overt open reading frames. Protist. 1999; 150**:**61-70.

- Willhoeft U., Tannich E. The electrophoretic karyotype of *Entamoeba histolytica*. Mol Biochem Parasitol. 1999; 99(1):41-53.

- Winckler T., Dingermann T. and Glockner G. *Dictyostelium* mobile elements: strategies to amplify in a compact genome. Cell Mol Life Sci. 2002; 59:2097-2111.

- Woodcock D.M., Lawler C.B., Linsenmeyer M.E., Doherty J.P. and Warren W.D. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. J Biol Chem. 1997; 272(12):7810–6.

- Xie H., Wang M., Bonaldo M. F., Smith C., Rajaram V., Goldman S. *et al.* Highthroughput sequence-based epigenomic analysis of Alu repeats in human cerebellum. Nucleic Acids Res. 2009; 37(13):4331-40.

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

- Yadav V.P, Mandal P.K., Bhattacharya A., Bhattacharya S. Recombinant SINEs are formed at high frequency during induced retrotransposition in vivo. Nat Commun. 2012; 3:854.

- Yadav V.P., Mandal P.K., Rao D.N., Bhattacharya S. Characterization of the restriction enzyme-like endonuclease encoded by the *Entamoeba histolytica* non-long terminal repeat retrotransposon EhLINE1. FEBS J. 2009; 276(23):7070-82.

- Yang D., Lu H., Erickson J.W. Evidence that processed small dsRNAs may mediate sequence-specific mRNA degradation during RNAi in *Drosophila* embryos. Curr Biol. 2000; 10(19):1191-200.

- Yang J., Malik H.S., Eickbush T.H. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements**.** *Proc Natl Acad Sci USA.* 1999; 96:7847-7852.

- Yang J., Malik H.S. and Eickbush T.H. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc. Natl. Acad. Sci. USA.1999; 96: 7847-7852.

- Yoder J.A., Walsh C.P., Bestor T.H. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 1997; 13:335–340.

- Young C.L., Britton Z.T., Robinson A.S. Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. Biotechnol J. 2012; 7(5):620-34.

- Zamudio N., Bourc'his D. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity. 2010; 105:92-104.

- Ziller M. J. *et al*. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013; 500:477–481.

Appendix

Primer Sequences

7.1 Primers used for Ion torrent sequencing

Primer Name	Sequence in 5'-3' direction
LINE1Frag1 FP	AAGAAAAGCAAACAAGACATAGAAAT
LINE1Frag1 RP	ATTCTTTCTTGTATCTCTTTTATTGTTA
LINE1Frag2 FP	GATAGATGTAAAATTGCAAGAATAATAAAT
LINE1Frag2 FP	TTCTTCTGAGATGGCTTGTTCTTCT
LINE1Frag3 FP	AAATAAATGAGATAGAAGGAAAGAAAATCA
LINE1Frag3 RP	AGATTTGTTTTTCTTTATCTCTTATTTC
LINE1Frag4 FP	GATGAAATTAAAGAAATCCTAAAGAAAAT
LINE1Frag4 RP	CAAACATGATTTTATTTGTATTGATTCT
LINE1Frag5 FP	GAAAAGGATGTCATCATCAAGAGAA
LINE1Frag5 RP	GTAAATACTTTCATAATTACATTTGTACA
LINE1Frag7 FP	CTATTGTGGCTGATCACAATATTAAT
LINE1Frag7 RP	CGATGTCAATTTCAATTTTTATTTCTTTAATTA
LINE1Frag1 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAAAAGCAAACAAGACATAGAAAT
LINE1Frag1 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGATATTCTTTCTTGTATCTCTTTTATTGTTA
LINE1Frag1 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATAAGAAAAGCAAACAAGACATAGAAAT
LINE1Frag1 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGATTCTTTCTTGTATCTCTTTTATTGTTA
LINE1Frag2 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGGATAGATGTAAAATTGCAAGAATAATAAAT
LINE1Frag2 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGATTTCTTCTGAGATGGCTTGTTCTTCT
LINE1Frag2 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATGATAGATGTAAAATTGCAAGAATAATAAAT
LINE1Frag2 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGTTCTTCTGAGATGGCTTGTTCTTCT
LINE1Frag3 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAATAAATGAGATAGAAGGAAAGAAAATCA
LINE1Frag3 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGAT AGATTTGTTTTTCTTTATCTCTTATTTC
LINE1Frag3 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATAAATAAATGAGATAGAAGGAAAGAAAATCA
LINE1Frag3 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGAGATTTGTTTTTCTTTATCTCTTATTTC
LINE1Frag4 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGGATGAAATTAAAGAAATCCTAAAGAAAAT
LINE1Frag4 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGATCAAACATGATTTTATTTGTATTGATTCT
LINE1Frag4 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATGATGAAATTAAAGAAATCCTAAAGAAAAT
LINE1Frag4 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGCAAACATGATTTTATTTGTATTGATTCT
LINE1Frag5 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGGAAAAGGATGTCATCATCAAGAGAA
LINE1Frag5 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGAT GTAAATACTTTCATAATTACATTTGTACA
LINE1Frag5 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATGAAAAGGATGTCATCATCAAGAGAA
LINE1Frag5 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGGTAAATACTTTCATAATTACATTTGTACA
LINE1Frag7 set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTATTGTGGCTGATCACAATATTAAT
LINE1Frag7 set1RP Adapter	CCTCTCTATGGGCAGTCGGTGAT CGATGTCAATTTCAATT-TTTATTTCTTTAATTA
LINE1Frag7 set2FP Adapter	CCTCTCTATGGGCAGTCGGTGATCTATTGTGGCTGATCACAATATTAAT
LINE1Frag7 set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGCGATGTCAATTTCAATT-TTTATTTCTTTAATTA
SINE1 for deep seq. FP	GCTGCAAAGGGTGCAGCAAGA
SINE1 for deep seq. RP	CCTTTGTTTGTTTTCTACCTTAATTTT
SINE1 Frag1set1FP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGGCTGCAAAGGGTGCAGCAAGA
SINE1 Frag1set1RP Adapter	CCTCTCTATGGGCAGTCGGTGATCCTTTGTTTGTTTTCTACCTTAATTTT
SINE1 Frag1set2FP Adapter	CCTCTCTATGGGCAGTCGGTGAT GCTGCAAAGGGTGCAGCAAGA
SINE1 Frag1set2RP Adapter	CCATCTCATCCCTGCGTGTCTCCGACTCAGCCTTTGTTTGTTTTCTACCTTAATTTT

7.2 Primers used for methylation study

Primer Name	Sequence in 5'-3' direction
P-ORF1 F	GGCCGCGGAGATCCTTTTCCAATGCAGG
P-ORF1 R	GGTACCCGTTTGAATCTTTCTATTTTGTG
P-ORF2 F	GCCCTCGAGTTGAAGTGTTGTATTGTTTTGC
P-ORF2 R	GCCGGTACCGAATATCTTCAATCTAGTCATCAT
CF	GTAAAGAAAAGCAAACAAGACATAGAAAT
CR	CAAGAAGTATATTCAATTAAAAAGAAA
07F1 (BS^-)	TCTTTTATATATTTTTATTCTTTTTATATTATTTGTT
07R1 (BS^-)	TTATTGTACTTATCCTGTTTCTTATTC
07F2 (BS^-)	GTTGTTGTTTCTATTTCTTTCTTCTATTATTTATCAA
07R2 (BS^-)	CTTCTTTGTCTTTGTTGTTTTTTAAACCCTT
07F1 (BS^+)	TTTTTTATATATTTTTATTTTTTTTATATTATTTGTT
07R1 (BS^+)	TTATTATACTTATCCTATTTCTTATTC
07F2 (BS^+)	GTTGTTGTTTTTATTTTTTTTTTTTATTATTTATTAA
07R2 (BS^+)	CTTCTTTATCTTTATTATTTTTTAAACCCTT
92F1 (BS^-)	AAACATATACCAATAATTTCATTTTTGATTTTTGAG
92R1 (BS^-)	GTTATTTGTATAATCTTTATTGTTATTGTATTTACC
92F2 (BS^-)	TTCATTTTTGATTTTTGAGTCTTGTTTTTAG
92R2 (BS^-)	TTATTGTATTTACCTTGTTTTTTATTTGTTTTTC
92F1 (BS^+)	AAATATATATTAATAATTTTATTTTTGATTTTTGAG
92R1 (BS^+)	ATTATTTATATAATCTTTATTATTATTATATTTACC
92F2 (BS^+)	TTTATTTTTGATTTTTGAGTTTTGTTTTTAG
92R2 (BS^+)	TTATTATATTTACCTTATTTTTTATTTATTTTTC
Amp1F (BS^+)	GGTTTTGTAAAGAAAAGTAAATAAGAT
Amp1R (BS^+)	CCTTATTTTTTATTTATTTTTCTTTATATCTTTC
Amp2F (BS^+)	AAATGAGATATAAGAAAGAATAAAAAAAATAAT
Amp2R (BS^+)	TTAATATCTTTTTCCTCCCAATCTTTTA
Amp1F (BS^-)	GGTCTTGTAAAGAAAAGCAAACAAGAC
Amp1R (BS^-)	CCTTGTTTTTTATTTGTTTTTCTTTGTGTCTTTC
Amp2F (BS^-)	AAATGAGATACAAGAAAGAATAAAAAAAATAAC
Amp2R (BS^-)	TTAATGTCTTTTTCCTCCCAATCTTTTG
CR1 (BS^-)	TCTTCTAAACCCTTTTTCCTTGGTGTATC
CR2 (BS^-)	CCTTGTTTTTTATTTGTTTTTCTTTGTGTCTTTC
CR3 (BS^-)	TTACCTTCATATTTTTTAAGTATTTCTTC
CR4 (BS^-)	ATCTTTTGTTTTTTCTTTAAACTCTTCTTTC
CR1a (BS^-)	ATGTATCGTTTACCTTTATTATTATAATT
CR2a (BS^-)	TCTTTCGTCTTCTAAACCCTTTTTCCTT
CR1 (BS^+)	TCTTCTAAACCCTTTTTCCTTAATATATC
CR2 (BS^+)	CCTTATTTTTTATTTATTTTTCTTTATATCTTTC
CR3 (BS^+)	TTACCTTCATATTTTTTAAATATTTCTTC
CR4 (BS^+)	ATCTTTTATTTTTTCTTTAAACTCTTCTTTC
HspR1	CAGCAACACGTCATCTTATAAC
HspR2	TTGTATTTCACTATTCAGCAACAC
qHspF	AGTCCAACCAATTTTCACTAAGCTCTATC
qHspR	GAATCCATTTGGCATTCCTCCTGG

7.3 Other primers sequences

Primer Name	Sequence in 5'-3' direction
ORF1dT PCRFP1	AACAAGAGAAGAATTAGACAACACAC
ORF1dT PCRFP2	TAGAGAAGAAGAAAACGATGACAC
ORF2dT PCR FP	GAAACAAAAATAGAAGAAATAATAAATGAAGG
U3sno FP	TAGACCGTACTCTTAGGATCATTTCT
U3sno RP	ATAGTCAGACACCCTAACATCACCTCTTG
Oligo dT 45mer	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
PEORF2RP3	GATTAATAAAGTTCTTCCACTTACCATG
EhSINE1For	GGCACGAGGGCACGTCTGAAACACC
EhSINE1Rev	TAAAAAGAAAAAAGTAATTAATTAAGTATT
HJ67-F	GCGGGTACCATGCAGGAAATACAACAAATT
HJ67-R	TTCTGTAATTTCTTCTTCAAATTCTTT
Cnst Spacer FP (Northern probe spacer)	AATACAACAGAACCAACAAATGGAATTTTAATTGAAGTGTTG
Cnst Spacer RP (Northern probe spacer	GATAGAGGTGAATAGTCATTGAAATATTCTTCATATC
LT64 FP (Northern probe A)	GCGGTACCATTTCAATGACTATTCACCTCTAT
LT64 RP (Northern probe A)	AGTGTCTATTCCTGGTGCTTTCCAGTT
BK49 FP (Northern probe B)	AACTGGAAAGCACCAGGAATAGACACT
BK49 RP (Northern probe B)	GAATATCTTCAATCTAGTCATCAT
DY32 FP (Northern probe C)	ATGATGACTAGATTGAAGATA
DY32 RP (Northern probe C)	TTTGGAGTTTCCGGCACCAATATATTTTGCCTTTCTTAA
DX11 FP (Northern probe D)	GGTGCCGGAAACTCCAAAGATAGGTTATATGTCCCTCTAGAA
DX11 RP (Northern probe D)	GCGAATTCCATGCTTTAACTGTAGTAGTTTT
EN SINE1 minus FP	ACATGACCATATAGGCATTATAATATGG
EN SINE1 minus RP	TCTACGGAGTATGTTTGGTTGTGATC
Spacer Luc FP	GCCCTCGAGTTGAAGTGTTGTATTGTTTTGC
ORF2 Luc RP	GCCGGTACCGAATATCTTCAATCTAGTCATCAT
SINE1 F (SINE 600)	GGCACGAGGGCACGTCTGAAACACC
EhSINE1 rev (SINE 600)	TAAAAAGAAAAAAGTAATTAATTAAGTATT
T7SINE1 F (*in vitro* SINE1 600 bp)	TAATACGACTCACTATAGGGAGAGGCACGAGGGCACGTC
T7SINEMidF (*in vitro* SINE1 120bp)	TAATACGACTCACTATAGGGAGACAAAGAGATTACTCCTTT
EhSINE1 MidF (120 bp)	GAGACCCACGCTCACCGGCGTAGTAATAAAATAATTCCT
RT(t) ACC for (RT FP)	GCGGTACCATGAATTATCGTCCTATCAG
RT(t) Bam rev (RT RP)	GCGGATCCTTATGCTTTAACTGTAGTAG

Publication

1. **Agrahari M.**, Gaurav AK., Bhattacharya A., Bhattacharya S. Cytosine DNA methylation at promoter of non LTR Retrotransposons and heat shock protein gene (Hsp70) of Entamoeba histolytica and lack of correlation with transcription status. **Mol Biochem Parasitol**. 2017; 212:21-27.

2. Gaurav AK., Kumar J., **Agrahari M.**, Bhattacharya A., Yadav VP., Bhattacharya S. 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**. 2016; 211:84-93.

Molecular & Biochemical Parasitology 212 (2017) 21–27

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Cytosine DNA methylation at promoter of non LTR retrotransposon and heat shock protein gene (HSP70) of *Entamoeba histolytica* and lack of correlation with transcription status

Mridula Agrahari [a], Amit Kumar Gaurav [a], Alok Bhattacharya [b], Sudha Bhattacharya [a,*]

[a] School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, India
[b] School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Article history:
Received 17 November 2016
Received in revised form
26 December 2016
Accepted 5 January 2017
Available online 9 January 2017

Keywords:
Non LTR retrotransposon EhLINE1
EhHsp70
Cytosine DNA methylation
Entamoeba histolytica
Transcriptional regulation
Bisulfite mapping

ABSTRACT

Non LTR retrotransposons (EhLINEs and EhSINEs) occupy 11% of the *Entamoeba histolytica* genome. Since promoter DNA methylation at cytosines has been correlated with transcriptional silencing of transposable elements in model organisms we checked whether this was the case in EhLINE1. We located promoter activity in a 841 bp fragment at 5′-end of this element by luciferase reporter assay. From RNAseq and RT-PCR analyses we selected a transcriptionally active and silent copy to study cytosine DNA methylation of the promoter region by bisulfite sequencing. None of the cytosines were methylated in either copy. Further, we looked at methylation status of a few selected cytosines in all 5′-intact EhLINE1 copies by single nucleotide incorporation opposite cytosine in bisulfite-treated DNA, where dGTP would be incorporated if the cytosine was methylated. Again we did not find evidence of cytosine methylation, indicating that expression status of this element was not correlated with promoter DNA methylation. To test for any role of cytosine methylation in transcriptional regulation of the *E. histolytica* Hsp70 gene in which the promoter is fully methylated under normal growth conditions, we checked methylation status and found that the promoter remained fully methylated during heat-shock as well, although transcription was greatly enhanced by heat-shock, showing that cytosine methylation is not a repressive mark for EhHsp70. Our data present direct evidence that promoter methylation, a common mode of transposon silencing, is unlikely to be involved in transcriptional regulation of EhLINE1, and reinforce the conclusion that promoter DNA methylation may not be a major contributor to transcriptional regulation in *E. histolytica*.

1. Introduction

Transposable elements have been extremely successful genome colonizers, with retrotransposable elements occupying as much as 40% of the sequenced mammalian genomes [1,2]. Since their continued movement through the genome would be potentially mutagenic, organisms have evolved ways to functionally silence these elements to maintain genome stability [3]. Intact copies of transposable elements with constitutive promoters are controlled by factors like DNA modification, chromatin environment, transcription factor availability and post-transcriptional regulation [3,4]. In mammals a principle mechanism for retrotransposon silencing in both germ cells and somatic cells is transcriptional repression through DNA methylation on cytosine residues in the context of CpG dinucleotides. Inhibition of DNA methyltransferases leads to increased expression of retrotransposons and endogenous retroviruses [5]. Transposon promoters are inactive when methylated, and suppression of their expression appears to be a primary function of cytosine methylation [6,7].

The genome of the early-branching protozoan parasite *Entamoeba histolytica* is rich in retrotransposons, with 11% of the genome being occupied by non long terminal repeat (LTR) retrotransposons. These belong to two major classes- the autonomous long interspersed nuclear elements (LINEs) and their short, non autonomous partners (SINEs) [8,9]. EhLINEs and EhSINEs are classified into three related families, of which EhLINE1 and EhSINE1 is the most abundant. EhLINE1 is 4.8 kb and codes for ORF1 which

* Corresponding author at: Lab 111, School of Environmental Sciences, Jawaharlal Nehru University, New Mehrauli Road, New Delhi, 110067, India.
E-mail address: sbjnu110@gmail.com (S. Bhattacharya).

http://dx.doi.org/10.1016/j.molbiopara.2017.01.001
0166-6851/

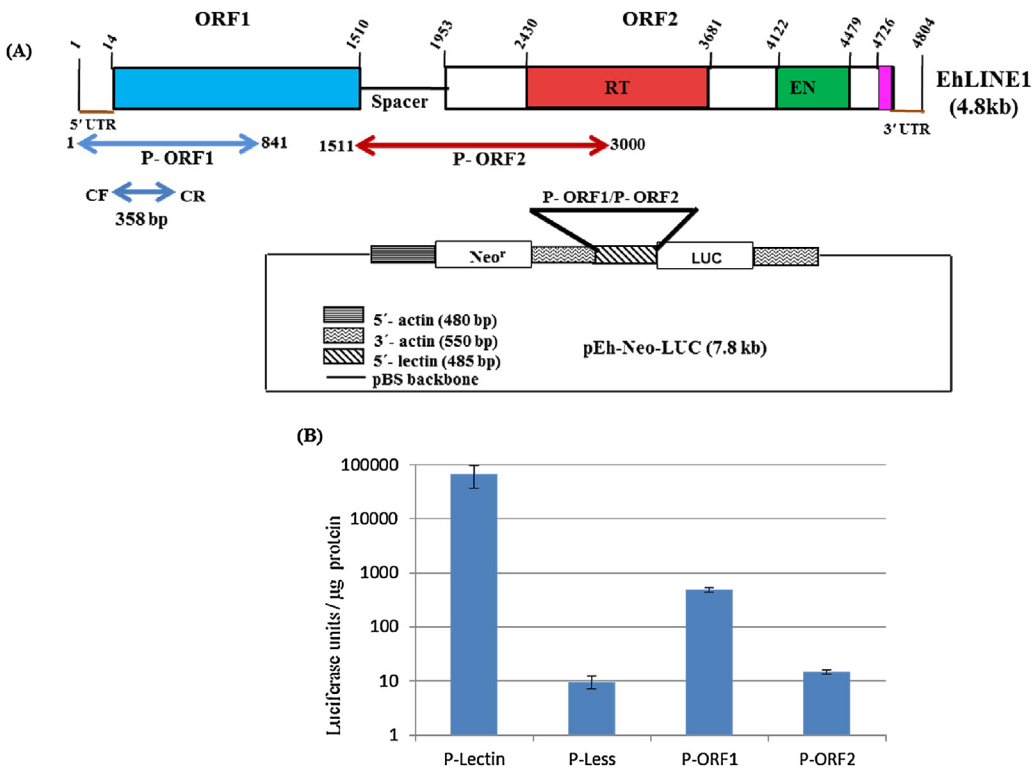


Fig. 1. Promoter analysis of EhLINE1 using luciferase reporter system, and identification of expressed EhLINE1 copies. (A) Schematic representation of EhLINE1 showing positions of the two ORFs and functional domains [8]. The fragments used for promoter analysis (P-ORF1 and P-ORF2), and site of insertion (replacing 5′-lectin) in pEhNeoLuc vector upstream to the luciferase reporter gene are shown. Also shown is the 358 bp region amplified by primers CF and CR from 5′-end of EhLINE1 to identify the expressed copies using RT-PCR, followed by cloning and sequencing (details in the text). (B) For promoter analysis cells were transfected with the indicated constructs, and luciferase activity was measured. P-lectin and P-less are the positive and negative controls, respectively. The data is an average of three independent experiments. (All primers used for PCR are listed in Supplementary Table S1).

is a nucleic acid binding protein, and ORF2 which has domains for reverse transcriptase and restriction enzyme-like endonuclease [10,11] (Fig. 1). Some copies of this element are transcriptionally active in *E. histolytica* since ORF1p could be detected in these cells. However, no ORF2p could be detected and active retrotransposition could not be demonstrated. Nevertheless, when ORF2p was ectopically over expressed in a *E. histolytica* cell line, the retrotransposition of EhSINE1 could be demonstrated [11].

The presence of cytosine methylation in *E. histolytica* DNA has been shown by using antibodies against 5-methyl cytosine. Methylated DNA was affinity purified using these antibodies and the predominant DNA class obtained was ribosomal DNA. The non LTR retrotransposon sequence of *E. histolytica* was also recovered from this affinity purification, pointing to the possibility that some cytosine residues in EhLINEs may be methylated [12,13]. The effect of DNA methylation on transcription activity was checked by 5-azacytidine (azaC) treatment to inhibit DNA methyl transferase. This study showed that cytosine methyaltion does regulate transcripton in *E. histolytica*, although of a limited set of genes [14]. The transcription of EhLINEs/SINEs could not be addressed in this study as the microarray did not contain probes corresponding to them. It is not known whether the transcription of these elements in *E. histolytica* is modulated by DNA methylation.

Since promoter DNA methylation has been correlated with transcriptional silencing in model systems [3] we wished to check whether the 5′-end of EhLINE1 (where the internal promoter is located in LINE elements) showed cytosine methylation and whether this correlated with transcription status of individual EhLINE1 copies. We also extended this analysis to the *E. histolytica* HSP70 gene which is known to be methylated at all cytosines in its promoter region [15] in normal cells, but the methylation status upon heat shock is unknown.

2. Materials and methods

2.1. Luciferase reporter assay for promoter mapping

Cloned EhLINE1 DNA [8] was used as template for PCR amplification of P-ORF1 and P-ORF2 fragments, cloned in pEhNeoLuc vector upstream to the luciferase reporter gene. Stably transfected cell lines were obtained and luciferase enzyme assay was performed as described [16]. *E. histolytica* culture and growth conditions were as described [17]. For transfection we used the method as described [18]. All primers used for PCR are listed in Supplementary Table S1.

2.2. RNA sequencing and data analysis

Total RNA was purified from exponentially growing *E. histolytica* cells using TRIzol reagent (Invitrogen) and 10 μg of it was used for paired end deep sequencing using Illumina HiSeq 2500 (v3 Chemistry) platform. From Paired-end reads unwanted sequences were removed including non-polyA tailed RNAs using bowtie2 (version 2.2.2) and in-house Perl scripts. About 35 million reads were obtained and on an average, ~90.13% of total reads passed > = 30 Phred score. The reads were aligned to the *E. histolytica* (HM1:IMSS) LINE1 reference sequences, downloaded from AmoebaDB, using RSEM v1.2.31 with default parameters and commands: "rsem-prepare-reference" and "rsem-calculate-expression". In the RSEM output the gene level expression was shown by read count, TPM and FPKM. Genes were considered expressed if read count was >10, and thus expressed and silent copies of EhLINE1 were categorized.

2.3. Bisulfite treatment, and DNA sequencing

E. histolytica genomic DNA was isolated using Wizard genomic DNA isolation kit, (Promega) and bisulfite (BS) modification of the DNA was done using EZ DNA Methylation-Lightning Kit, (Zymo) according to the manufacturer's instructions. Bisulfite treated (or untreated) DNA was PCR amplified in 50ul reaction containing 0.25 mM each dNTP, 2U zymoTaq DNA polymerase along with 1uM BS-converted (or normal) primers/nested primers. Cycling conditions were: 95 °C/10 min, 40 cycles of 95 °C/30 s, 50–55 °C/30 s, 72 °C/30–60 s subsequently followed by 72 °C/7 min. Amplified products were cloned in pGEMT-EASY vector (Promega) and subjected to Sanger sequencing.

2.4. Single nucleotide incorporation

Primers were labeled at 5′end by polynucleotide kinase (NEB) with 40 μCi of [γ-^{32}P]-ATP. Bisulfite treated/untreated DNA (200 ng) was mixed with 50 μM dGTP/dATP, phusion buffer (1×), labeled primer (~60,000 counts) and 1U phusion polymerase (NEB). After heating at 95 °C/5 min, annealing was done at 50 °C/2 min, followed by incorporation at 72 °C/10 min. The product was denatured at 95 °C/5 min followed by snap chill on ice/5 min and separated on 7M- 6% Urea-PAGE. Incorporation was detected by Phosphor Imager (Typhoon FLA 9500, GE).

2.5. Quantitative real time-PCR

DNase I treated total RNA (400 ng) was reverse transcribed in a 30 μl reaction using primer qHspR, with Revertaid-RT (Fermentas). PCR was performed in 7500 Real Time PCR System (Applied Biosystems) using SYBR green PCR Master Mix, 2 pmol of primers qHspF, qHspR and 2 μl of cDNA (1:4 dilution). Actin (control) was amplified in parallel. The conditions were: annealing at 50 °C/10 s followed by 1 cycle of denaturation at 95 °C/10 min and 40 cycles at 95 °C/10 s and 60 °C for 1 min. Cycle threshold values (C_t) were analyzed by the SDS1.4 software (Applied Biosystems) and all samples were analyzed in triplicates in three independent experiments. Reactions without cDNA were used as no template control and no RT controls were also set up to rule out genomic DNA contamination. The expression values are relative to the levels of actin.

2.6. RNA isolation and northern hybridization

Total RNA from approximately 5 × 106 cells was purified using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. For northern analysis RNA samples (20 μg) were resolved in 1% formaldehyde agarose gel in buffer [0.1 M MOPS (pH 7.0), 40 mM sodium acetate, 5 mM EDTA (pH 8.0)] and 37% formaldehyde at 4 V/cm. The RNA was transferred on to GeneScreen plus (Perkin Elmer) nylon membranes. Alpha-P32 dATP-labeled probes were prepared by random priming method using Decalabel DNA labeling kit (Thermo scientific). Hybridization and washing conditions for RNA blots were as per manufacturer's protocol.

3. Results

3.1. Eh LINE1 promoter is located at its 5′-end

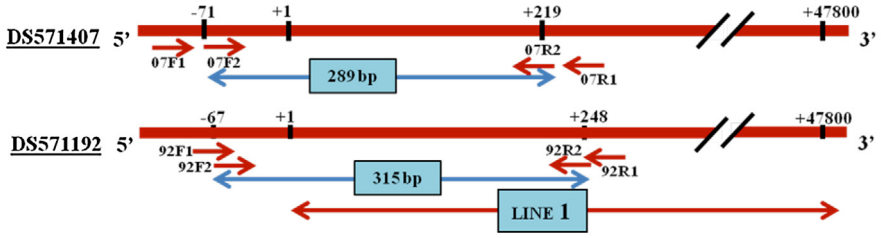
Transcription of LINEs generally initiates from the internal promoter at the 5′-end of the element to produce a single transcript covering the entire LINE sequence [19]. We were interested to know whether cytosine methylation of promoter sequences, a known repressive mechanism in many organisms, determined the expression status of individual EhLINE1 copies. For this we first demonstrated promoter activity at the 5′-end of EhLINE1 by using luciferase reporter system. A copy of EhLINE1 (located on scaffold DS571192) that showed high expression from RNA seq data (details in the next section) was selected for this analysis. A fragment of 841 bp from the 5′-end of this copy was cloned upstream of luciferase in pEhNeo-Luc vector [16] to obtain the construct P-ORF1 (Fig. 1A). As controls we used the E. histolytica lectin promoter (P-lectin), and a construct with no promoter (P-less). We also used a construct (P-ORF2) containing a 1.5 kb sequence downstream of ORF1 to discount the possibility of a second promoter. The results showed that there was robust expression of luciferase in cells transfected with the P-ORF1 construct, while the expression with P-ORF2 and P-less constructs was at background levels (Fig. 1A).

3.2. Cytosine methylation status of the promoter region of transcriptionally active and silent EhLINE1 copies

To select transcriptionally active or silent copies, the expression status of individual EhLINE1 copies was estimated from RNA-seq analysis. RNA sequence data were obtained by paired-end deep sequencing using Illumina platform (details in the "Materials and Methods" section). Of the 35 million reads, >90% aligned with E. histolytica genome, and the sequencing was done with two independent biological replicates. (Details of the complete RNA-seq analysis will be published elsewhere). Reads were aligned to EhLINE1 reference sequences and a copy was considered expressed if the read count was >10. To validate the data from RNA-seq we cloned the expressed EhLINE1 sequences obtained by reverse transcriptase (RT)-PCR of total RNA. For this we made PCR primers (CF and CR shown in Fig. 1A) from conserved region of ORF1 5′-end (these would amplify a minimum of 92 EhLINE1 copies as judged from sequence identity), and used them for RT-PCR to obtain a 358 bp amplicon. This was cloned and the inserts were sequenced from 30 randomly picked colonies. From these data the EhLINE1 copy in scaffold DS571192 was selected for further analysis as an expressed copy, since it ranked amongst the top 10 EhLINE1-expressed sequences in RNA-seq; it was present in the 30 colonies sequenced from RT-PCR; and it showed the closest match (96% identity) with the consensus EhLINE1 sequence [8]. For silent copy we selected the EhLINE1 sequence in scaffold DS571407 since it showed zero expression in RNA-seq data, was not scored in the RT-PCR analysis (its sequence matched completely with the RT-PCR primers used), and was full-length. Methylation status of the 5′-end of these two copies was checked by bisulfite (BS) treatment of total genomic DNA, which converts unmethylated cytosines to uracil in DNA, while methylated cytosines are protected [20]. The individual copies were amplified using locus-specific upstream primers (Fig. 2). Nested primers were used to get specific amplicons. Sequence alignment of the silent and expressed copies is shown in Supplementary Fig. S1. Both amplicons had a total of 36 cytosines. BS-converted primers (C–T) only amplified BS-treated DNA, and *vice versa*, showing that the bisulfite treatment was successful. Amplicons were sequenced to determine the extent of cytosine methylation. The 289 bp amplicon from the silent copy contained 219 bp of EhLINE sequence with 2 CpG sites, while the 315 bp amplicon from the expressed copy contained 248 bp of LINE sequence, with 3 CpG sites. Sequence analysis showed that all the cytosine residues in DNA were converted to thymine upon treatment with bisulfite, showing that none of these cytosine residues were methylated in either of the two copies (Supplementary Fig. S1). To show that the absence of cytosine methylation was not due to a technical problem we introduced methyl residues at CpG sites by treating genomic DNA with CpG Methyltransferase (M.SssI), followed by bisulfite conversion and amplicon generation as described above. The expressed EhLINE1 copy in scaffold DS571192 has 3 CpG sites. Their methylation status was checked in M.SssI-treated DNA and 2 out of the 3 cytosines were protected from bisulfite conver-

(A)

S No.	Accession No.	LINE1 Position	Length	Expression status	Expected amplicon size	No. of Cs /CpG in amplicon
1	DS571407	14251-19035	4.78kb	Silent	289bp	36/2
2	DS571192	10681-15465	4.78kb	Expressed	315bp	36/3



(B)

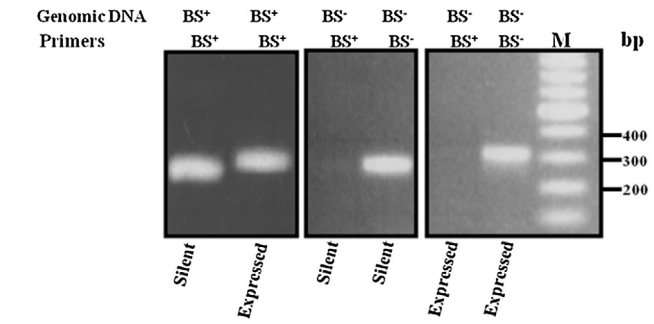


Fig. 2. Cytosine methylation status of the promoter of transcriptionally active and silent EhLINE1 copies. (A) The two expressed and silent EhLINE1 copies selected to check cytosine methylation status at their promoter site are shown. Bisulfite converted (BS^+) and normal (BS^-) primers, including nested primers (07F1, F2, R1, R2 series for silent copy and similar 92 series for expressed copy), for bisulfite PCR (BS-PCR) were designed from the locations shown. The forward primers were upstream of EhLINE1 so as to amplify the specific EhLINE1 copy. (B) BS-PCR of expressed and silent EhLINE1 copies with BS-treated and untreated genomic DNA to confirm the primer specificity. The BS+ primers amplified only the BS-treated DNA. Amplicon sequences are given in Supplementary Fig. S1.

sion, showing that the observed absence of cytosine methylation in genomic EhLINE1 copies was not likely to be an experimental artefact. (All 3 cytosines may not be protected due to incomplete methylation by M.SssI).

3.3. *Detection of cytosine methylation at selected sites in a larger subset of EhLINE1 copies*

In the above experiment we looked at methylation of all cytosines in the 5′-regions of only two EhLINE1 copies. We next determined the methylation of a few selected cytosines but in a larger subset of EhLINE1 copies. For this we adopted the following approach. We aligned all (256) 5′-intact copies of EhLINE1 and looked for conserved CpG sites in the first 500 bp. Four such sites were found to occur frequently, with at least one of the four sites present in 160 copies. Primers were designed from conserved sequences flanking these sites to obtain two amplicons containing two CpG sites each (Fig. 3A). The amplicons were obtained from bisulfite-treated genomic DNA using bisulfite-converted primers as described above. For each CpG site a complementary primer was used with its 3′ ending at C (complementary to the G residue in CpG). Primers were also designed to score three non-CpG sites in amplicon 1. These were present in a larger number of copies, with at least one site present in 200 copies. The end-labeled primers were annealed with the amplicons from bisulfite-treated DNA and allowed to incorporate a single nucleotide (A or G), which would reflect the methylation status of the cytosine at that site. Amplicons from non-bisulfite treated DNA obtained with non bisulfite-converted primers were used as control. The data showed that dATP was incorporated with bisulfite-treated DNA, while dGTP was incorporated with untreated DNA at all seven sites, showing lack of extensive cytosine methylation at these sites (Fig. 3B). If a small subset of the copies were methylated, their number could be estimated by determining the ratio of radioactivity at the +1 position compared with origin in the dGTP lane, by densitometry. This ratio was close to zero for dGTP in all samples, showing negligible levels of methylation.

3.4. *The promoter of* E. histolytica *HSP70 gene remains methylated during heat shock when transcription is up regulated*

Our data with EhLINE1 copies showed negligible cytosine DNA methylation at the sites examined by us in both expressed and silent copies, indicating that DNA methylation is unlikely to be involved in transcriptional regulation of these elements. One of the genes whose promoter is fully methylated in normal *E. histolytica* cells is HSP70 (Fisher et al., 2006). The methylation status of this gene has not been checked during heat stress, which could directly correlate transcriptional control of this gene with promoter methylation. We used the methods described above for EhLINEs to check the methylation of HSP70 promoter region (−201 to +42) both under normal and heat-stressed (42 °C for 60 min) conditions. Sequencing of bisulfite-treated DNA showed that all cytosines were methylated in both conditions (Supplementary Fig. S2). Further we also checked methylation of two selected CpG sites by incorporation of dGTP/dATP, which again showed that both sites were methylated in normal and heat-stressed cells (Fig. 4). As a control, the EhLINE1 showed no methylation in the same DNA samples. We looked at expression status of HSP70 in heat-shocked cells by northern hybridization. Transcript levels were negligible in nor-

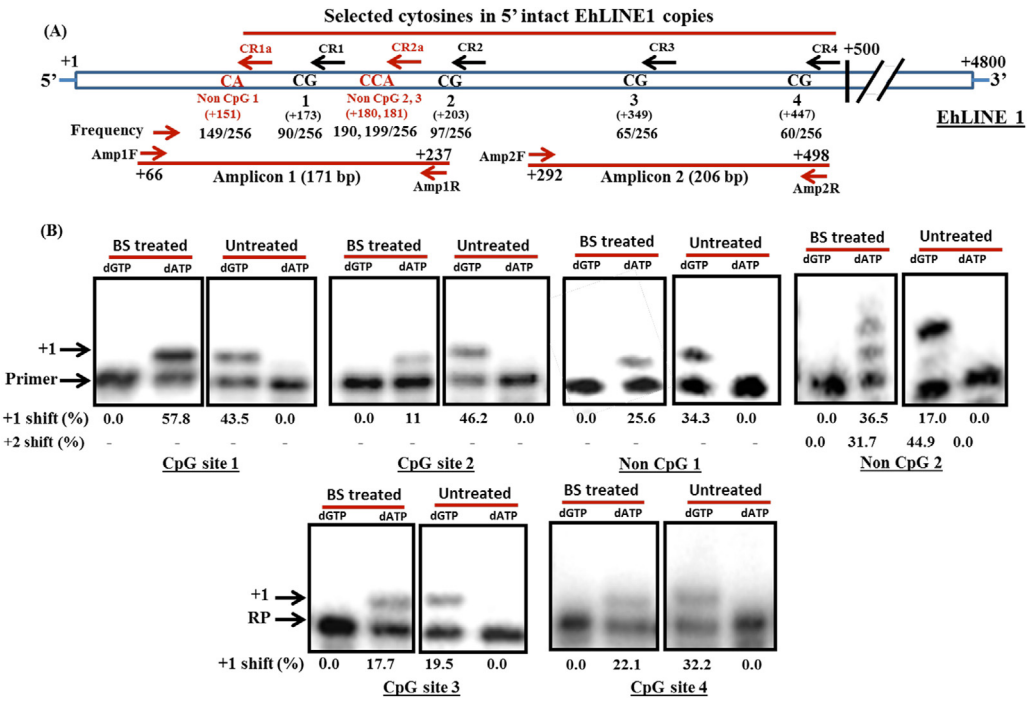


Fig. 3. Detection of cytosine methylation at selected sites in the promoter of a large subset of EhLINE1 copies. (A) The location of four conserved CpG sites and three Non CpG sites in the 5′–500 bp region of EhLINE1 copies are shown, along with positions of the two amplicons containing these sites. Of the 256 5′-intact EhLINE1 copies the number of copies in which the selected CpG residues occur is indicated below each CpG site. Reverse primers CR1, 2, 3 & 4 (shown above each site) were designed such that they end at the 'G' residue, to be used for single nucleotide incorporation assay opposite the 'C'. Non CpG Reverse primers CR1a and CR2a were also designed with the same strategy. (B) Single nucleotide incorporation assay. Amplicons 1 and 2 were obtained from BS-treated and untreated DNA using primer pairs Amp 1F/1R and Amp 2F/2R respectively). Amplicon DNAs were annealed with respective end-labeled reverse primer for each site and extended in presence of either dGTP or dATP. +1 is the shift after nucleotide incorporation and percent +1 and +2 shift measured by densitometry is indicated.

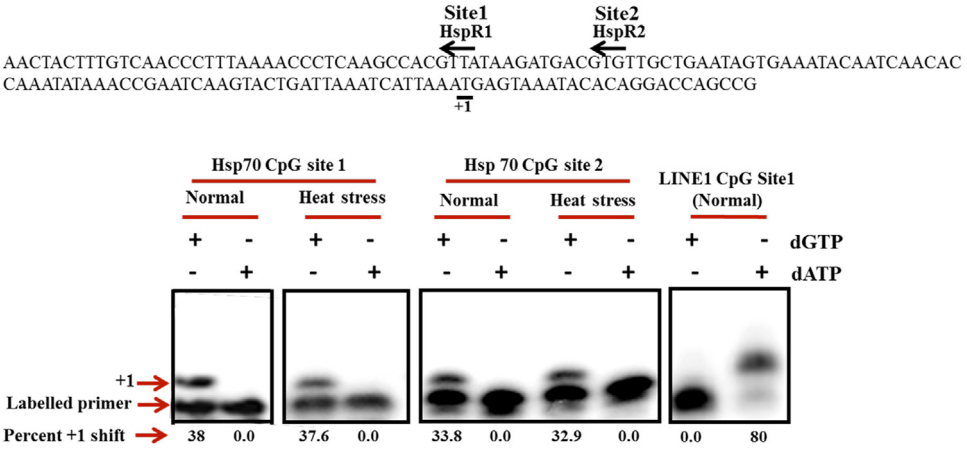


Fig. 4. Single nucleotide incorporation assay in the Hsp70 gene copy (EAL45068) at selected CpG sites: Two CpG sites were selected in the region known to be methylated [15] and primers ending at 'G' were made. DNA obtained from normal and heat-stressed cells (42 °C/60 min) was treated with bisulfite, annealed with primers, and the assay was done as described for EhLINE1 in Fig. 3. In the same assay, DNA from non heat-stressed cells was used for EhLINE1 CpG site 1, which showed opposite results compared with Hsp70.

mal cells and as expected transcription increased to high levels upon heat shock (Fig. 5). *E. histolytica* has 17 copies of HSP70 gene reported in the data base. To specifically determine transcript levels of the gene copy that we used for cytosine methylation analysis we used gene-specific primers for quantitative RT-PCR, which showed 8.5.-fold upregulation of this copy upon heat shock. Since the cytosine methylation status of the HSP70 gene promoter remained unchanged although its transcription increased tremendously, the data directly demonstrate that DNA methylation was not involved in transcriptional regulation of this gene.

4. Discussion

Cytosine DNA methylation at promoter regions is a common mode of retrotransposon silencing in a variety of organisms [21]. Earlier studies in *E. histolytica* have indicated the possibility of methylation of EhLINE sequences since antibodies against the *E. histolytica*-methylated LINE binding protein (EhMLBP) interacted with EhLINE sequences *in vivo* as shown by chromatin immunoprecipitation [22]. The protein also bound to another highly repetitive DNA- the rDNA of *E. histolytica*, suggesting that it could have a role in modulating the expression of highly repetitive DNA. However,

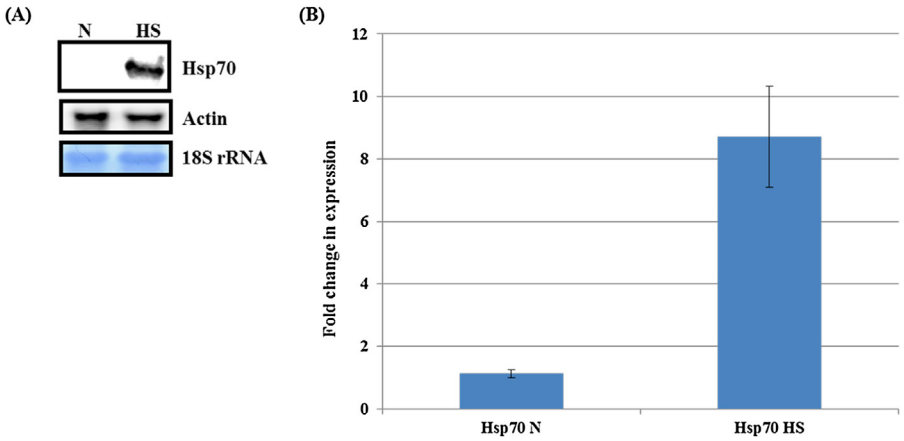


Fig. 5. Expression of Hsp70 gene in normal (N) and heat-stressed (HS) conditions. (A) Northern blot analysis with Hsp70 probe in N and HS conditions. Actin was used as a control. (B) Expression level of the Hsp70 copy (EAL45068) used for DNA methylation analysis, quantified by qRT-PCR with primers qHspF and qHspR showed 8.5 fold increase in transcript levels in HS cells.

direct demonstration of promoter methylation and transcription attenuation of these sequences has not been done. In a study to see the effect of azaC on *E. histolytica* gene expression, the transcription status of EhLINEs/SINEs could not be scored as these sequences were absent in the microarray [14]. Hence we undertook this study to determine the level of cytosine methylation at EhLINE1 promoter and its correlation with transcriptional repression.

Bisulfite sequencing of a 200 bp region at the 5′-end of an expressed and silent EhLINE1 copy showed complete lack of cytosine DNA methylation in both copies. To confirm that this was not due to a technical problem we showed that the Hsp70 promoter was fully methylated in our cells, as previously reported (Fisher et al., 2006). Since EhLINE1 is present in 967 copies it is possible that some of these copies may be methylated, and the two copies analyzed by us (Fig. 2) were exceptions. We used the strategy of single nucleotide incorporation opposite cytosine in bisulfite treated DNA to check the methylation status of selected cytosines in a larger subset of EhLINE1 (62.5% and 78.5% of the 5′-intact copies contained CpG or non CpG cytosines, respectively). Again, we did not find any cytosine methylation, whereas all the cytosines in Hsp70 were scored as methylated by this method also.

Our data show that EhLINE1 promoter sequences are almost devoid of cytosine DNA methylation. In an earlier report, Harony et al. [13] also could not demonstrate cytosine methylation of EhLINE, although they were studying the RT sequence and not the promoter. It is possible that the transcriptional status of EhLINEs may be regulated by other mechanisms which remain to be explored. For example, histone methylation instead of DNA methylation might suppress transcription of these elements, as reported for some mammalian SINE sequences [23]. DNA methylation also seems to be absent in *Schizosaccharomyces pombe* and *Caenorhabditis elegans* where epigenetic silencing is mediated by histone modifications [24,25]. Retrotransposon transcript levels are also regulated at the level of degradation, as these RNAs are known to be specifically targeted for degradation by small RNAs, mainly the PIWI-interacting RNAs [26]. The major class of small RNAs identified in *E. histolytica* are Argonaute-associated, 27 nt long RNAs which predominantly map to the open reading frames of protein-coding genes, with only ~7% mapping to EhLINE/SINE sequences [27]. These antisense small RNAs are involved in gene silencing [28] however, their role, if any, in EhLINE/SINE silencing has not been studied.

Our study adds to earlier reports which indicated that DNA methylation has a rather limited effect on transcription attenuation in *E. histolytica* [14,15]. In the study by Ali et al. [14], transcription of only 2.1% genes was significantly modulated by 5-azaC treatment. The authors looked for any association of genes in the vicinity of EhLINEs/SINEs with modulation by 5-azaC and did not find any correlation, which fits with our data showing lack of cytosine methylation in EhLINE1. The study of Fisher et al. [15] looked at phenotypic changes in cells overexpressing Ehmeth, a methyltransferase of the Dnmt2 family. These cells showed pleiotropic changes (multinucleation, resistance to oxidative stress), and the transcription of HSP70 gene was upregulated. Since this gene is fully methylated even in normal cells, its methylation status is not expected to change in the Ehmeth-overexpressed cells, and the observed upregulation could be an indirect effect. We provide direct evidence that the *E. histolytica* HSP70 promoter DNA remains methylated during heat shock when the gene is actively transcribed. Thus, cytosine methylation is not a repressive mark for this gene.

Our data show for the first time that DNA cytosine methylation of promoter, which is a common mechanism of transposon silencing in a variety of organisms, is unlikely to modulate the transcription of EhLINE1. *E. histolytica* shares similarities with *Dictyostelium discoideum* in which only a small fraction (~0.2%) of the genome is methylated and the only methytransferase known is a homolog of Dnmt2 (DnmA). In DnmA knockout cells the LTR retrotransposon Skipper was upregulated but the expression of another retrotransposon DIRS-1 remained unaffected [29], suggesting only a limited regulatory role of DNA methylation in retrotransposon expression. Interestingly, a recent report shows that Dnmt2, a tRNA methylase could efficiently methylate cytosines in DNA in the context of a covalent DNA-tRNA hybrid [30], suggesting that this enzyme could methylate DNA *in vivo* under specific contexts. The role of this enzyme in cytosine DNA methylation in *E. histolytica*, and the physiological conditions under which it may be activated need to be understood. The global mechanism responsible for transcriptional regulation of the large number of EhLINE copies in *E. histolytica* remains to be discovered.

Acknowledgements

M.A. and A.K.G. are recipient of Senior Research Fellowship from CSIR and ICMR respectively. S.B. received research grant from DST and DBT.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molbiopara.2017.01.001.

References

[1] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[2] R.H. Waterston, Mouse Genome Sequencing Consortium: initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[3] J.H. Crichton, Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline, Cell. Mol. Life Sci. 71 (9) (2014) 1581–1605.
[4] Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development, Nature 447 (2007) 425–432.
[5] R. Ollinger, Meiosis and retrotransposon silencing during germ cell development in mice, Differentiation 79 (3) (2010) 147–158.
[6] J.A. Yoder, Cytosine methylation and the ecology of intragenomic parasites, Trends Genet. 13 (1997) 335–340.
[7] J.A. Hackett, Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline, Development 139 (2012) 3623–3632.
[8] A.A. Bakre, et al., The LINEs and SINEs of Entamoeba histolytica: comparative analysis and genomic distribution, Exp. Parasitol. 110 (2005) 207–213.
[9] H. Lorenzi, Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species, BMC Genomics 9 (2008) 595.
[10] P.K. Mandal, An *Entamoeba histolytica* LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease, Eukaryot. Cell 3 (2004) 170–179.
[11] V.P. Yadav, et al., Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*, Nat. Commun. 22 (3) (2012) 854.
[12] O. Fisher, et al., Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase (Ehmeth) in the protozoan parasite *Entamoeba histolytica*, Nucleic Acids Res. 32 (1) (2004) 287–297 (9).
[13] H. Harony, et al., DNA methylation and targeting of LINE retrotransposons in *Entamoeba histolytica* and *Entamoeba invadens*, Mol. Biochem. Parasitol. 147 (2006) 55–63.
[14] I.K. Ali, Growth of the protozoan parasite *Entamoeba histolytica* in 5-azacytidine has limited effects on parasitegene expression, BMC Genomics 8 (2007) 7.
[15] O. Fisher, Pleiotropic phenotype in *Entamoeba histolytica* overexpressing DNA methyltransferase (Ehmeth), Mol. Biochem. Parasitol. 147 (2006) 48–54.
[16] S. Shrimal, et al., Serum-dependent selective expression of ehTMKB1-9, a member of *Entamoeba histolytica* B1 family of transmembrane kinases, PLoS Pathog. 6 (6) (2010) e1000929, 3.
[17] L.S. Diamond, A new medium for the axenic cultivation of *Entamoeba histolytica* and other *Entamoeba*, Trans. R. Soc. Trop. Med. Hyg. 72 (4) (1978) 431–432.
[18] N. Sahoo, Calcium binding protein of *Entamoeba histolytica*, Arch. Med. Res. 31 (2000) S57–59.
[19] G.D. Swergold, Identification: characterization and cell specificity of a human LINE- 1 promoter, Mol. Cell. Biol. 10 (12) (1990) 6718–6729.
[20] M.F. Fraga, M. Esteller, DNA methylation: a profile of methods and applications, BioTechniques 33 (2002) 632–649.
[21] J.A. Law, Establishing, maintaining and modifying DNA methylation patterns in plants and animals, Nat. Rev. Genet. 11 (2010) 204–220.
[22] T. Lavi, Sensing DNA methylation in the protozoan parasite *Entamoeba histolytica*, Mol. Microbiol. 62 (2006) 1373–1386.
[23] R.A. Elbarbary, et al., Retrotransposons as regulators of gene expression, Science 351 (6274) (2016) aac7247, 12.
[24] I.M. Hall, Establishment and maintenance of a heterochromatin domain, Science 297 (5590) (2002) 2232–2237.
[25] T.A. Volpe, Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi, Science 297 (5588) (2002) 1833–1837.
[26] M. Reuter, Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing, Nature 480 (2011) 264–267.
[27] H. Zhang, Small RNA pyrosequencing in the protozoan parasite Entamoeba histolytica reveals strain-specific smallRNAs that target virulence genes, BMC Genomics 14 (2013) 53.
[28] L. Morf, et al., Robust gene silencing mediated by antisense small RNAs in the pathogenic protist Entamoeba histolytica, Nucleic Acids Res. 20 (2013) 9424–9437.
[29] M. Kuhlmann, Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi, Nucleic Acids Res. 33 (19) (2005) 6405–6417.
[30] Steffen Kaiser, et al., The RNA methyltransferase Dnmt2 methylates DNA in the structural context of a tRNA, RNA Biol. 0 (2016) 1–11.

Molecular & Biochemical Parasitology 211 (2017) 84–93

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Research Paper

Functionally conserved RNA-binding and protein-protein interaction properties of LINE-ORF1p in an ancient clade of non-LTR retrotransposons of *Entamoeba histolytica*

Amit Kumar Gaurav[a], Jitender Kumar[a], Mridula Agrahari[a], Alok Bhattacharya[b], Vijay Pal Yadav[a], Sudha Bhattacharya[a,*]

[a] School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, India
[b] School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Article history:
Received 13 June 2016
Received in revised form
17 November 2016
Accepted 24 November 2016
Available online 25 November 2016

Keywords:
Entamoeba histolytica
RNA Binding Protein
Retrotransposon
EhLINE1
R2 group

ABSTRACT

Retrotransposons are mobile genetic elements found in most organisms. Their origin and evolution is not very well understood. Retrotransposons that lack long terminal repeats (non-LTR) have been classified based on their reverse transcriptase (RT) and endonuclease sequences into groups, of which R2 is the most ancient. Its members contain a single open reading frame (ORF) while there are two ORFs in the other groups, of which ORF2 contains the RT and endonuclease sequences. It is thought that ORF1 was added later to the single-ORF-containing elements, and codes for a protein with nucleic acid binding activity. We have examined the non-LTR retrotransposons in *Entamoeba histolytica*, an early-branching parasitic protist, which belongs to the R2 group. However, unlike other members of R2, *E. histolytica* contains two ORFs. Here we show that EhLINE1-ORF1p is functionally related to the ORF1p found in the non-R2 groups. Its N-terminal region has RNA-binding activity and its C-terminal has a coiled coil domain which participates in protein-protein interaction. It lacks sequence-specificity of RNA-binding and binds to EhLINE1-RNA fragment and ribosomal RNA with comparable affinities. Our study suggests that ORF1p could have evolved independently to maintain functional conservation.

1. Introduction

Retrotransposons lacking long terminal repeats (non LTR) are present in a large variety of organisms, including humans where they occupy about one-third of the genome [1–3]. They consist of autonomous elements that encode the functions required for retrotransposition and non-autonomous elements, which are respectively called long or short interspersed nuclear elements (LINEs or SINEs). A generic LINE element may be 5–7 kb and may contain a single ORF or two ORFs. They are broadly classified into five groups [4,5]. Elements with a single ORF belong to the more ancient R2 group and they contain a restriction enzyme-like (REL) endonuclease, while elements of the other groups code for apurinic/apyrimidinic endonuclease (APE) [6]. The R2 group consists of several members from protists, and some members from insects, nematodes, fish and green algae [4,5,7,8]. The C-terminal region of elements with a single ORF is similar to ORF2 (in elements with two ORFs) and has well-conserved domains corresponding to reverse transcriptase (RT) and DNA endonuclease, which have both been used for phylogenetic analysis [4,9]. The sequence of the N-terminal part of elements with a single ORF (or ORF1 in the other groups), however, is much less conserved. In spite of sequence heterogeneity ORF1 of most non-LTR retrotransposons serves the conserved functions of nucleic acid binding, and ribonucleoprotein (RNP) formation with the LINE transcript, which is essential for retrotransposition [10–14].

About 11% of the genome of the early-branching protist *Entamoeba histolytica* is composed of non-LTR retrotransposons called EhLINEs and EhSINEs [15–18]. Full-length EhLINE1 is 4.8 kb, and although it belongs to the R2 group, it contains two ORFs [15,19]. ORF1 is 1494 bp, with no known functional domains. ORF2 is 3093 bp and contains conserved RT and REL endonuclease domains. We have earlier functionally characterized the endonuclease [20,21], and have also demonstrated *in vivo* retrotransposition in ORF2 over-expressing *E. histolytica* cell lines [19]. However, nothing is known about the properties of ORF1p from EhLINE1.

* Corresponding author at: Lab 111, School of Environmental Sciences, Jawaharlal Nehru University, New Mehrauli Road, New Delhi, 110067, India.
E-mail address: sbjnu110@gmail.com (S. Bhattacharya).

http://dx.doi.org/10.1016/j.molbiopara.2016.11.004
0166-6851/

The ORF1p in various groups of non-LTR retrotransposons contains several distinct domains specific to each clade. The human L1 ORF1p possesses three structural domains, namely, coiled-coil (CC) needed for homotrimerization, a non canonical RNA recognition motif (RRM), and C-terminal domain (CTD) located in a basic region, which facilitate formation of L1-RNP [22]. In contrast the only identifiable domain in the ORF1p of zebrafish LINE is an esterase domain [14]. Evidence of nucleic acid binding property of ORF1p was first demonstrated using human L1RNA [23], and has subsequently been demonstrated with the polypeptide from mouse [11–13,24], zebrafish [14], and *Drosophila* [25]. ORF1p also possesses nucleic acid chaperone and self-interaction properties [13,14,24].

Amongst the elements in the R2 group, the best studied is R2Bm from *Bombyx mori*, which inserts in an extremely sequence-specific manner in the 28S rRNA gene [26]. This encodes a single 1113 amino acid polypeptide with a central RT domain, a downstream REL endonuclease domain, and upstream DNA-binding region with Zn finger motifs, and Myb domain [27]. Two molecules of the protein bind symmetrically to the two ends of the target DNA site, and this requires presence of either the 3′-end or the 5′-end of the R2Bm RNA transcript. Direct binding of protein to DNA has been shown by EMSA, while binding to RNA has been inferred from the requirement for RNA in the DNA binding. Conserved amino acid residues immediately upstream of the RT domain are required for RNA binding [28]. The functional role of RNA in the retrotransposition reaction has been well-defined by these studies. However, the physical interaction of the protein with R2 RNA and formation of ribonucleoprotein has not been demonstrated.

Since EhLINE1 belongs phylogenetically to the R2 group but contains two ORFs we were interested to know whether its ORF1 is functionally similar to the N-terminal domain of the R2Bm protein in terms of binding to DNA, or whether it has the properties of ORF1p in elements with two ORFs. In the present study we have tested the biochemical properties of EhLINE1 ORF1p and demonstrated that it binds preferentially to single stranded RNA through its N-terminal domain, while the C-terminal domain is involved in self-interaction. This resembles the properties of ORF1p in lineages with two ORFs, although the relative location of RNA binding and coiled coil domains in mammalian L1ORF1p is reverse of that in EhLINE1 ORF1p. To our knowledge this is the first functional study of ORF1p encoded by a parasitic protist of the R2 group.

2. Materials and methods

2.1. DNA sequence of ORF1

GSS (genome sequence survey) clones ENTHJ67 (AZ684953) and ENTEK39 (AZ542852) having maximum similarity with consensus sequence of ORF1 of EhLINE1 [15] and lacking any stop codons were selected for reconstruction of ORF1. The DNA was used in an overlapping PCR to reconstruct full-length ORF1. Sub-fragments of ORF1p were generated by PCR using primers from desired locations and they, including full-length ORF1p, were cloned in His or GST-tagged expression vectors.

2.2. E. coli *strains*

E. coli strain DH5α was used for all recombinant DNA work. *E. coli* BL21 (Rosetta) was used for expression of recombinant ORF1p and its sub-fragments.

2.3. Expression and purification of recombinant proteins

His-tagged sub-fragments of ORF1 protein (N-ter 23, N-ter 34 and C-ter 37) were expressed and purified by Ni-NTA as described [20]. All GST-tagged proteins (ORF1p, C-ter 37 and N-ter 23) were expressed in BL21 (Rosetta) by cloning in pGEX4T-1 vector. *E. coli* cells were grown at 37 °C till OD_{600} 0.5 and induced with 0.5 mM IPTG and further grown at 18 °C for 6–9 h. The cells were harvested and protein was purified using glutathione sepharose 4 fast flow beads as per manufacturer's instructions (GE healthcare). Fractions containing purified protein were identified by SDS-PAGE and then pooled and dialyzed against dialysis buffer [50 mM Tris-Cl (pH 8), 100 mM NaCl, 5 mM DTT and 30% glycerol]. The purified protein was quantified and stored in aliquots at −80 °C.

2.4. In vitro *synthesis of RNA*

Radiolabeled SINE1-RNA or rRNA were *in vitro* transcribed as described [40] using RiboMax large Scale RNA Production System (Promega).

2.5. Electrophoretic mobility shift assays

P^{32} labeled *in vitro* transcribed RNA (0.4 ng or ∼14,000 cpm) or DNA (EhSINE1-RNA, rRNA, EhSINE1-ssDNA, EhSINE1-dsDNA) substrates were incubated with indicated amount of purified recombinant ORF1p at ice for 15 min in 15 μl of 1X EMSA buffer as described [39]. RNA-protein complexes were fractionated by electrophoresis (8 V/cm at 4 °C, 3 h) through 5% native polyacrylamide gels (1:50 bisacrylamide:acrylamide) with 1% glycerol in 0.5X TBE. Gels were dried and autoradiographed using phosphorimager. For competition assays indicated amount of unlabeled competitor was added to the reaction.

2.6. GST pull down assay

GST-ORF1p or GST protein was immobilized on glutathione sepharose 4 fast flow beads. 25 μl of ORF1p or GST bound beads were incubated with His-tagged ORF1 sub-fragments (N-ter 23, N-ter 34 and C-ter 37) at 4 °C with rotation in 300 μl of NET-N+ buffer as described [14]. Glutathione sepharose beads were washed 5 times with NET-N+ buffer followed by SDS-PAGE and western blotting.

2.6. Bioinformatic analysis

All sequences were extracted from NCBI database or Amoeba DB (http://amoebadb.org/amoeba/). pI of ORF1p from various organisms (as described in 'Results') or their fragments were determined using Expasy ProtParam tool (http://web.expasy.org/cgi-bin/protparam/protparam). Prediction of nucleic acids binding residues/motifs was done by using BindN (http://bioinfo.ggc.org/bindn/) [41] and PPrint (http://www.imtech.res.in/raghava/pprint/) [42]. Other tools including ScanProsite (http://prosite.expasy.org/scanprosite/) [43], HHpred (toolkit.tuebingen.mpg.de/hhpred) [31] and SMART tool (smart.embl-heidelberg.de) [44] were also tested for prediction of nucleic acids binding residues/motifs. For coiled coil prediction multiple tools were used including SMART, Marcoil (http://toolkit.tuebingen.mpg.de/marcoil), multicoil (http://groups.csail.mit.edu/cb/multicoil/cgi-bin/multicoil.cgi) [45] and coils/Pcoils (http://toolkit.tuebingen.mpg.de/pcoils) [46].

3. Results

3.1. Domain analysis of EhLINE1 ORF1p and comparison with other protists

We have earlier shown by nucleotide sequence analysis that EhLINEs potentially encode two ORFs, of which ORF1 lacks easily

recognizable functional domains [15]. The polypeptide corresponding to EhLINE1 ORF1p is expressed in *E. histolytica* cells [19]. To understand the functional properties of this polypeptide, we undertook a comparative analysis of EhLINE1 ORF1p with the sequences of LINE-encoded ORF1p from other organisms, notably protists.

Since ORF1p is putative nucleic acid binding protein, which are usually basic, the pI of different sub-regions of ORF1p has been used to get an indication of possible nucleic acid binding sites and draw a comparison of the polypeptide from different species [24]. We checked the pI of ORF1p (or the region upstream of RT in case of a single ORF) of LINE sequences from various parasitic protists. Of the protists tested, *E. histolytica*, *Entamoeba dispar*, *Entamoeba invadens*, *and Giardia lamblia* belong to different clades in the R2 group (all of which have the REL endonuclease); and *Naeglaria gruberi* contains the APE-endonuclease and does not belong to the R2 group. The R2Bm sequence from *Bombyx mori* (of the R2 group) was also analyzed, and mammalian L1 sequences were included for comparison. In all cases the pI of the selected polypeptide was basic (except GilM of *G. lamblia*) (Fig. 1). The pI of three separate regions was also checked. The distribution of basic, acidic and neutral segments varied amongst the species examined. These sequences were further searched for presence of identifiable functional domains using various tools described in 'Materials and Methods'. Using these tools we could only find significant stretches of coiled coil (CC) domains, which facilitate protein-protein interactions. Interestingly, we found easily identifiable CC domains in all the *Entamoeba* LINEs and in the *N. gruberi* Proto 1_5 sequence at their C-terminii. No CC domains could be found in the *Giardia* sequences (Fig. 1). In general there was a positive correlation between the region with basic pI and predicted nucleic acid binding (NB) regions. In *E. histolytica*, *E. dispar*, and all three elements of *N. gruberi* the N-terminal region which is the most basic, showed best prediction of RNA binding (using PPrint tool). The *E. invadens* protein showed high prediction of RNA binding in the middle region. This region of *G. lamblia* proteins also contains well conserved CCHH motif which may be involved in nucleic acid binding [9,29], although such a motif present in an acidic region may not bind nucleic acids [30]. Overall, this analysis showed several conserved features in the domain distribution of ORF1p of EhLINE1, EdLINE1 and the *N. gruberi* LINEs. The unusual features of the Giardia elements, especially GilM could be explained because this element is highly degenerate [29].

The R2Bm sequence was analyzed by PPrint and it showed NB domains at its N-terminus and C-terminus. No CC domain could be predicted with any of the tools used. In mammalian ORF1ps the nucleic acid binding domain was located in the C-terminal region, while a CC domain was present in the N-terminus which was the reverse of EhLINE1 ORF1p (Fig. 1).

3.2. Identification of functional domains in EhLINE1ORF1p

We used several tools to identify possible functional domains in EhLINE1 ORF1p (briefly described above). The BindN tool predicts RNA binding residues based on side chain pKa value, hydrophobicity index and molecular mass of amino acid. It is also capable of predicting DNA binding residues. Another tool PPrint was also used for the same prediction. Both tools found the same stretch of amino acids at the N-terminus of EhORF1p with high RNA binding probability (supplementary Fig. S1). BindN predicted poor DNA binding ability of EhORF1p.

The tool ScanProsite did not give any appreciable hits with EhLINE1 ORF1p, while myhits showed nuclear localization signal (NLS). This was further analyzed with NLS mapper which predicted two NLS stretches [a monopartite NLS (from 338 to 348 aa), and a bipartite NLS (120–142 aa)]. Both SMART and HHpred identified only a long coiled coil domain at the C-terminus (supplementary Fig. S1). This was further confirmed by MARCOIL which revealed a coiled coil stretch of 282 aa (position 146–428) having well identified heptads [32] with prediction stringency of ≥98%, and several stretches of 100% probability. 'Multicoil' revealed a coiled coil stretch of 214 aa (position 206–419) and oligomerization prediction showed higher probability of dimer formation as compared to trimer (supplementary Fig. S2). SMART tool and coils/Pcoils predicted a coiled coil stretch of 234 aa (191–424) and 225 aa (201–425) respectively. Together, the data from these predictions located RNA binding region (from amino acids 19–90, 117–125 and 268–280), coiled coil (146–428 aa) and nuclear localization signal (120–142, and 338–348 aa) in EhLINE1 ORF1p (Fig. 2 and supplementary Fig. S1).

3.3. EhLINE1 ORF1p binds to RNAs of different sequences

Full-length EhLINE1 ORF1p and its sub fragments (N-ter 23, N-ter 34, C-ter 37) were cloned in *E. coli* expression vectors with either His or GST tags (Fig. 2; the full view of blots is given in supplementary Fig. S3). While the N-ter 23 and C-ter 37 polypeptides were expressed from appropriate deletion constructs of the EhLINE1 ORF1, the N-ter 34 polypeptide was expressed from a full-length EhLINE1 ORF1 copy containing a stop codon at aa position 277. Affinity purified proteins (Fig. 2) were tested for nucleic acid binding ability by electrophoretic mobility shift assay (EMSA) with various types of nucleic acids.

GST-tagged full-length EhLINE1 ORF1p (termed ORF1p) was incubated with RNA (65 nt) taken from 3′-end of EhSINE1 (termed as SINE1-RNA) which is similar in sequence to the 3′ end of EhLINE1. Since the only sequence similarity between EhSINE1 and EhLINE1 lies in the 75 nt sequence at 3′ end, this conservation may be important for SINE mobilization [20]. P^{32}-labeled SINE1-RNA was prepared by *in vitro* transcription. To perform EMSA, molar excess of the ORF1p (18 nM) was incubated with P^{32}-labeled SINE1-RNA (1.3 nM) [33] and electrophoresed through native PAGE. A shift in the mobility of RNA was observed in the presence of ORF1p, indicating that ORF1p could interact with EhSINE1-RNA (Fig. 3). Under the same assay conditions a non specific protein (BSA) did not show any shift. In addition a GST-tagged sub-fragment of ORF1p (N-ter 23) purified using the same procedure did not show any shift (Fig. 4), indicating that the observed shift was not due to GST tag. *In vitro* studies of ORF1p from other non-LTR retrotransposons show their high affinities toward RNA as compared to ssDNA or dsDNA [24]. To test the binding capacity of EhORF1p with different nucleic acids, we performed EMSA with end-labeled ssDNA and dsDNA of the same 65 nt sequence from 3′-end of EhSINE1 as was used for SINE1-RNA. Poor interaction of ORF1p (18 nM) was detected with ssDNA (1.3 nM) or dsDNA (1.3 nM) (Fig. 3), even when ORF1p concentration was increased to 30 nM. This result is different from the data reported with mammalian L1 ORF1p which also binds most efficiently to RNA but does bind to ssDNA (and poorly to dsDNA) [24].

To test the binding capability of ORF1p with RNAs other than EhSINE1, an 80 nt RNA was *in vitro* transcribed from rDNA sequence of *E. histolytica* and used for EMSA. This RNA sequence, which showed no significant similarity with SINE1-RNA (Bioedit identity matrix score 0.26) interacted efficiently with ORF1p and gave a reduced mobility complex (Fig. 3). This shows that at least *in vitro*, under the conditions of our assay, ORF1p did not discriminate between RNA sequences.

3.4. RNA binding activity is located in the N-terminal part of EhORF1p

To locate the region of EhORF1p responsible for nucleic acid binding, we performed EMSA with all three fragments of EhORF1p,

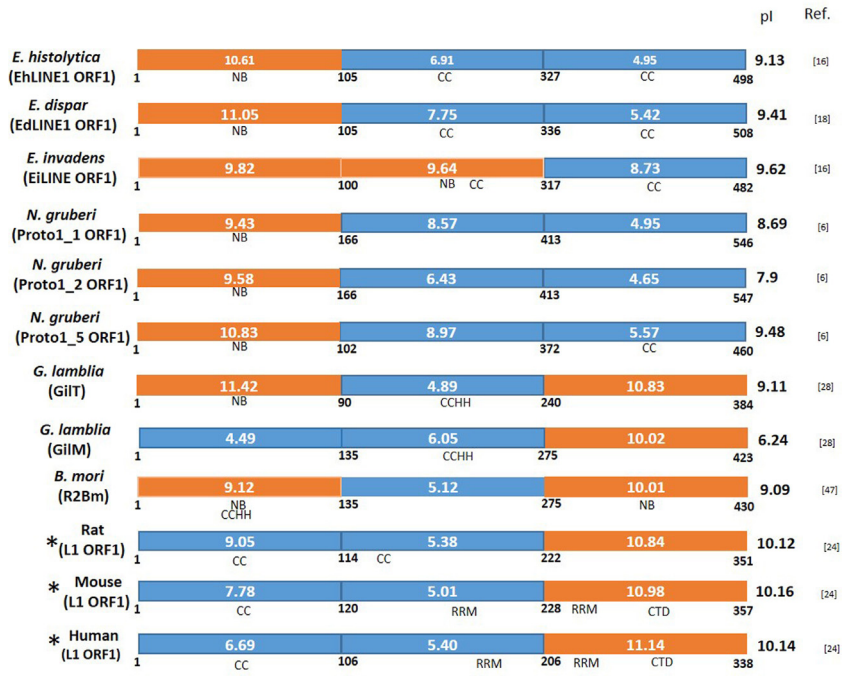


Fig 1. Comparison of amino acid sequence of ORF1p (or aa stretch upstream of RT) of LINEs from various organisms. The ORF1 amino acid sequence was divided into three segments based upon pI of each region [24]. The number below the junction of each box indicates amino acid position. The pI of each segment is written in the respective box, and total pI is written at the right of each element. Accession numbers of sequences compared are AZ684953 AZ542852 (EhLINE1 ORF1), AF313478.1 (EdLINE1), EU099446.1 (EiLINE1), AF433876.1 (GilT), AF433875.1 (GilM), T18197 (R2Bm), Q63303 (Rat L1ORF1), P11260 (Mouse L1ORF1), AH005269 (Human L1ORF1), *N. gruberi* Proto1_1, Proto1_2 and Proto 1_5 sequences were obtained from Repbase reports 2009 9(6). References (Ref) for source of all ORF1p sequences used in the analysis are given on the right. Sequences indicated with '*' were previously published [24]. Predicted functional domains are indicated. NB, nucleic acid binding; CC, coiled coil; RRM, RNA recognition motif; CTD, C-terminal domain [47].

N-ter 23, N-ter 34 and C-ter 37 (Fig. 4). Reactions containing 1.3 nM labeled SINE1-RNA were incubated with purified and dialysed polypeptides at the indicated concentrations for 15 min at 25 °C. Complexes were separated by 5% native PAGE and visualized by autoradiography (Fig. 4). No shift was observed with N-ter 23 (144 nM) and C-ter 37 (84 nM). However, with N-ter 34 we could see shift of radioactive SINE1-RNA starting from 10 nM input protein. As compared to full-length EhORF1p, N-ter34 showed less affinity for RNA as it failed to completely shift 1.3 nM SINE1-RNA even at 160 nM protein while full length ORF1p completely shifted it at 23 nM protein. We would like to add that since the two polypeptides had different tags (His-tag for N-ter 34 and GST-tag for full length ORF1p), this could also contribute to the observed difference in affinity. From this experiment we conclude that the first 277

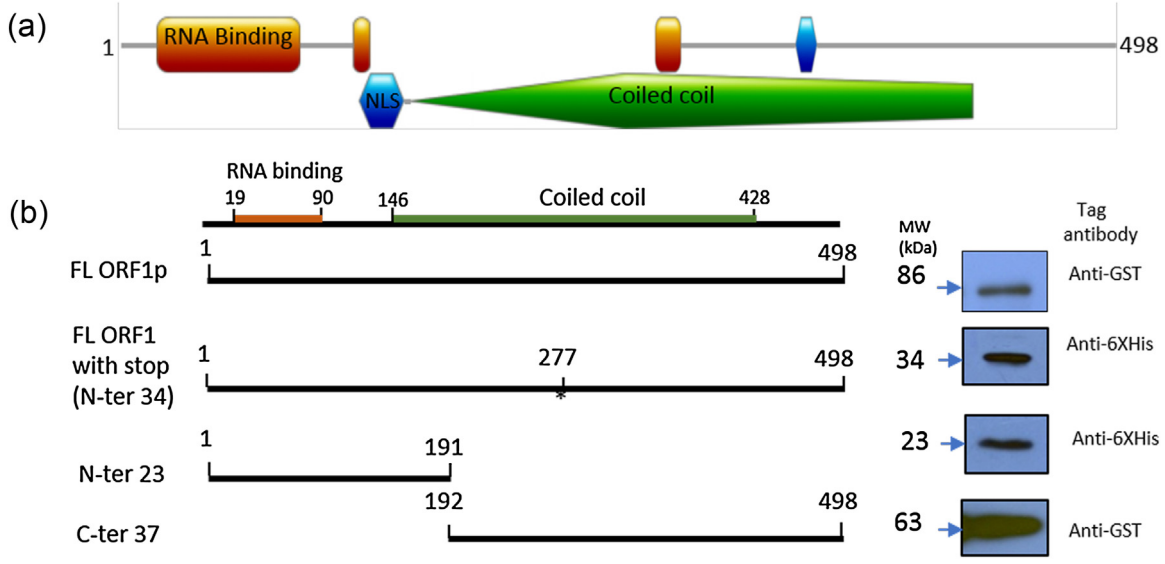


Fig. 2. ORF1p and its sub-fragments. (a). Schematic representation of domain structure of ORF1p showing stretches of RNA binding (19–90, 117–125 and 268–280 aa) and a coiled coil domain (146 aa to 428, as predicted by MARCOIL tool). Blue diamonds show predicted nuclear localisation signals (see text for details). (b). Full length ORF1p and sub-fragments were expressed and purified using different tags, as described in 'Materials and Methods'. The purified polypeptides were detected by western analysis using tag-specific antibodies.

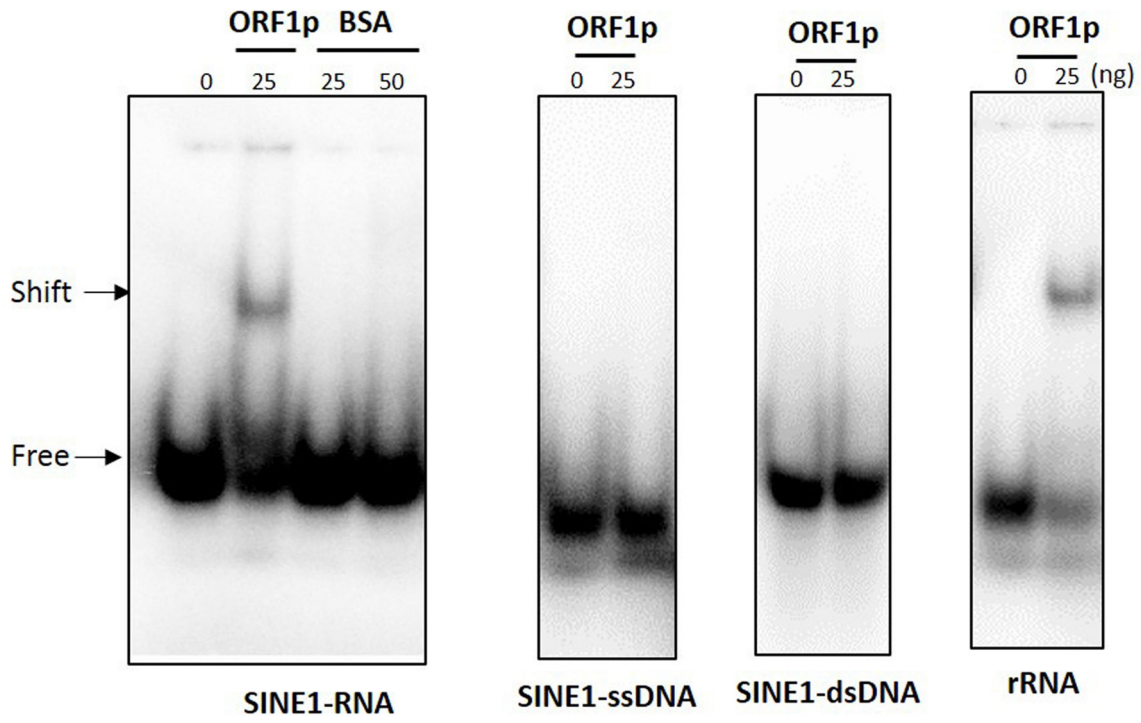


Fig. 3. Nucleic acid binding activity of EhORF1p. EMSA was performed with ORF1p and various nucleic acids. 65 nt EhSINE1-RNA (from 3′-end of EhSINE1) or 80 nt segment of rRNA was *in vitro* transcribed, and labeled with ^{32}P UTP. RNA (0.4 ng or ~14,000 cpm) was incubated with ORF1p. Protein *vs.* RNA molar ratio ranged from 0 to 13.8. BSA alone was used as a control. Electrophoresis was carried out through 5% native PAGE under cold conditions as mentioned in 'Materials and Methods' and gels were dried and autoradiographed by Phosphorimager. The DNA samples were 5′-end labeled for EMSA. High molecular weight complex (Shift) and Free probe (Free) are indicated by arrows.

aa of EhORF1p are sufficient for RNA binding, which corroborates with the prediction from sequence analysis.

3.5. ORF1p has comparable binding affinity to both SINE1-RNA and rRNA

To look for any quantitative difference in binding interaction of ORF1p with SINE1-RNA and rRNA, EMSA was performed with increasing concentration of ORF1p (0–23 nM) in presence of BSA. Complete shift of the free probe was observed at high concentration of ORF1p (Fig. 5a and b). The band intensities of bound and unbound RNA were determined, and the average values (three independent determinations) of fraction of bound RNA were plotted against the protein concentration (Fig. 5c). A linear regression was obtained by plotting the average of log bound/free {log (Y/(1-Y)), where Y is the fraction of bound RNA} against log concentration

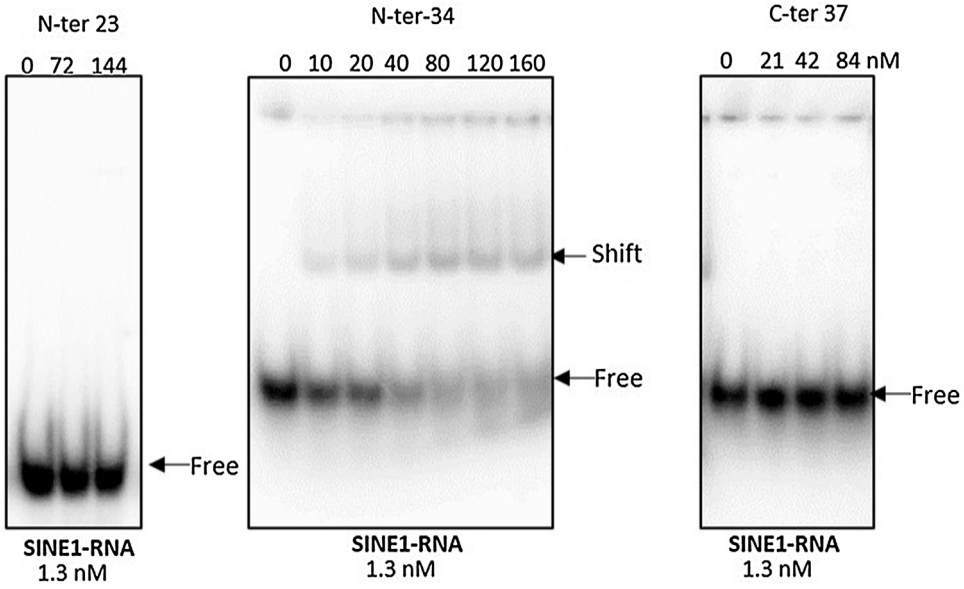


Fig. 4. RNA binding ability of various ORF1p sub-fragments: Recombinant N-ter 23 (GST- tagged), N-ter 34 (His-tagged) and C-ter 37 (GST- tagged) sub-fragments were used to test the RNA binding ability. In all EMSA experiments radiolabeled SINE1- RNA was used, as already described. C-ter 37 lacks RNA binding ability. Protein *vs.* RNA molar ratio ranged from 0 to 110.7 (N-ter 23), 0 to 123 (N-ter 34) and 0 to 64.6 (C-ter 37).

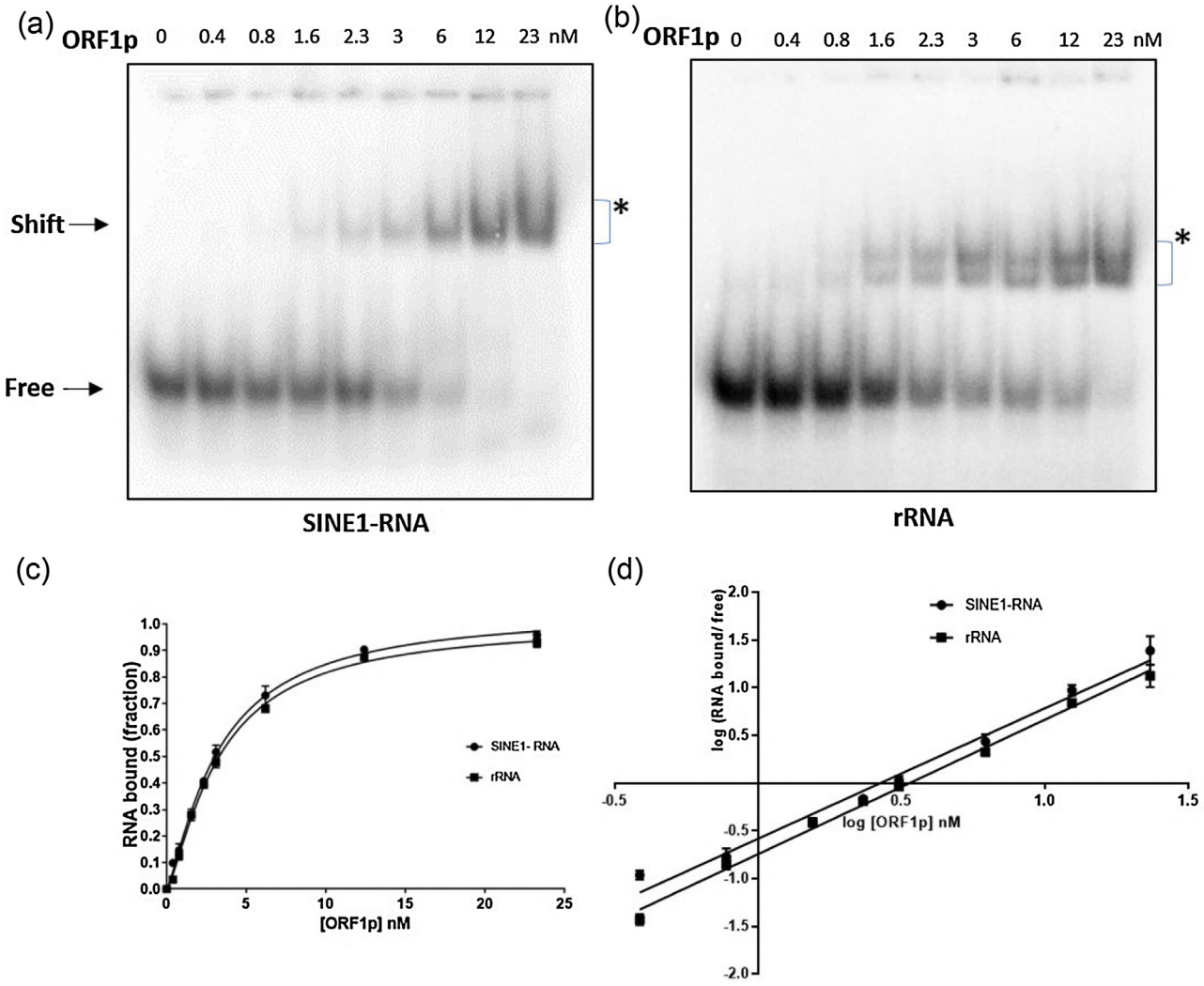


Fig. 5. Binding affinity of EhORF1p with SINE1-RNA and its comparison with rRNA. (a) & (b). EMSA with increasing concentration of ORF1p and SINE1-RNA (1.3 nM) or rRNA (1.3 nM). Protein *vs.* RNA molar ratio ranged from 0 to 17.7. EMSA conditions were similar as in Fig. 3. The values of 'Shift' and 'Free' fractions were determined by densitometry. (c). The average value (three independent determinations) of fraction of bound RNA was plotted against corresponding ORF1p concentration and fitted in a non-linear regression curve. (d). Log of bound/free was plotted against log of ORF1p concentration. Kd values were determined as described in the text.

of EhORF1p (Fig. 5d). K_d values, obtained from the x intercept, for rRNA and SINE1-RNA were calculated to be 3.36 ± 0.25 and 2.67 ± 0.15 nM respectively. Almost similar binding affinity with these two different types of RNA shows that EhLINE1 ORF1p may not recognize any particular RNA sequences under *in vitro* conditions. We checked the predicted folding pattern of the two RNA fragments using RNAfold (drawn with VARNA tool), and found that both RNAs formed stem-loop structures, with the SINE1-RNA forming a more compact structure compared with rRNA (supplementary Fig. S4). It remains to be seen whether secondary structure is important for the binding of ORF1p with RNA.

3.6. Binding specificity of EhORF1p evaluated by competition assays

The relative affinity of EhORF1p for different nucleic acids was also examined by competition assay with increasing concentration of unlabeled competitor (SINE1-RNA, rRNA, ssDNA or dsDNA). EhORF1p was used at a concentration of 6 nM which produces a moderately intense reduced mobility complex. Addition of unlabeled SINE1-RNA or unlabeled rRNA resulted in robust competition with few counts seen in the bound complex at high concentration of competitor (Fig. 6a). The average values corresponding to the fraction of bound radioactive complex as a function of competitor concentration were plotted (Fig. 6b). Both the SINE1- RNA and rRNA competed with the labeled probe to similar extent, and 80% of the probe was displaced at 5-fold competitor concentration. This data again suggests that binding affinity of EhORF1p is similar for both types of RNA. In contrast to RNA, unlabeled ssDNA and dsDNA were not able to displace labeled RNA from the bound complex (Fig. 6a). We tested these competitors at a very high concentration and found that only 20% displacement was achieved at concentrations up to 100-fold. This further strengthens our previous observations that EhORF1p interacts poorly with DNA.

3.7. EhLINE1 ORF1p protein-protein interaction is mediated through its C-terminal region

Evident from the sequence analysis, ORF1p contains a long coiled coil domain at the C-terminal. Major part of it lies in the C-ter 37 sub fragment, while N-ter 34 contains a small stretch and N-ter 23 lacks any significant coiled coil region. We tested all three sub fragments for interaction with full length ORF1p. GST pull-down assay was done with GST-tagged full length ORF1p and

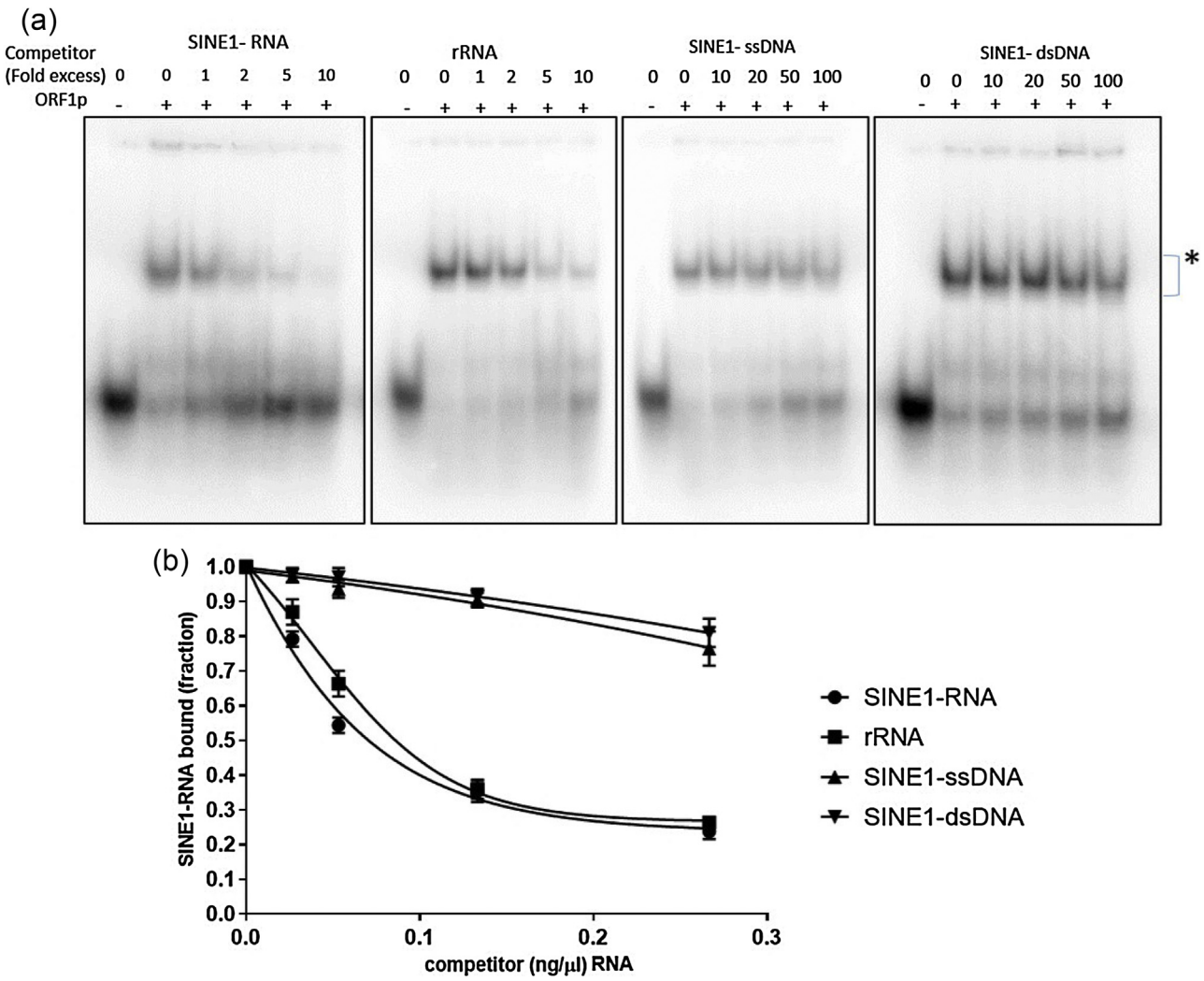


Fig. 6. Competition assays of EhORF1p binding with various nucleic acids. (a). The competition assays were carried out by incubating 6 nM of EhORF1p with radiolabeled SINE1-RNA (1.3 nM) in the presence of varying concentration of unlabeled nucleic acids as indicated. Protein *vs.* labelled RNA molar ratio was 4.6 (b). Plot of bound SINE1-RNA fraction (average of three independent determinations) against competitor concentration. Standard deviation is indicated as error bars (±SD).

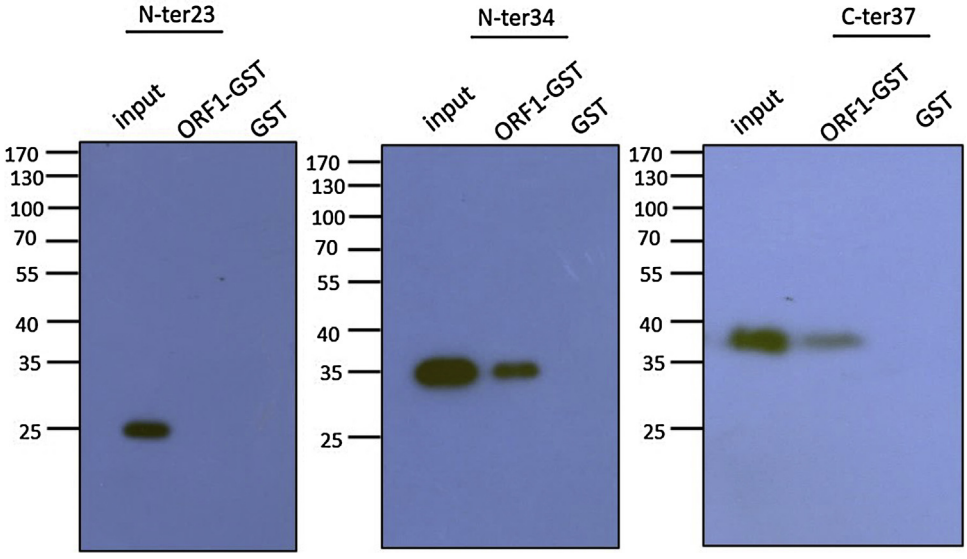


Fig. 7. Protein-protein interaction domain lies at the C-terminus. Interaction of full length GST-tagged ORF1p with its sub fragments was studied by incubating it with 6X His tagged sub fragments (N-ter 23, N-ter 34 and C-ter 37). The sub fragments were incubated with glutathione sepharose bound ORF1p at 4 °C overnight with gentle rotation. Glutathione sepharose beads were subsequently washed, boiled with SDS loading buffer and electrophoresed. Western blotting was performed with anti- His antibody. Among the three sub-fragments, only two (N-ter 34 and C-ter 37) were able to interact with full length EhORF1p. GST protein alone was used as negative control (GST).

6XHis tagged sub-fragments. Equal amount of 6X His tagged proteins were taken as input (Fig. 7). The data showed that N-ter 34 and C-ter 37 both interacted with EhORF1p. N-ter 23 sub-fragment showed no interaction. GST protein alone was used as negative control. These data corroborate with the predicted location of coiled coil domain in EhLINE1 ORF1.

4. Discussion

Phylogenetic classification of non-LTR retrotransposons has been done based on the conserved sequences of their reverse transcriptase and endonuclease domains [4]. This analysis has shown that elements with a single ORF generally belong to the more ancient R2 group, and their endonuclease is of the REL-type, while the later branching elements have two ORFs and they have an APE endonuclease. The *E. histolytica* EhLINE1 belongs to the R4 clade in the R2 group. Its ORF2 C-terminal has the REL endonuclease, and contains the highly conserved CCHC, $PDX_{12-14}D$, RHD, and KXXXY sequence motifs [15,20,34]. Unlike other members of the R2 group, many copies of EhLINE1 contain two ORFs (due to presence of stop codon between ORF1 and 2), while some may also contain a single ORF (in which stop codon is missing), as determined earlier from sequence analysis [15]. Polypeptides corresponding in size to the full-length EhLINE1 ORF or ORF2 could not be detected in *E. histolytica*, while a 60 kDa polypeptide corresponding to ORF1 was detected by western analysis, suggesting that ORF1p is expressed independently, and constitutively, in *E. histolytica*. Thus EhLINE does not strictly obey the generalization from earlier analysis that all members of the R2 group contain a single ORF, and it is possible that as more genomes are analyzed, the *E. histolytica*-type situation may be encountered more frequently.

Well-conserved domains could not be found in EhLINE ORF1, which is also the case for LINEs from other organisms. Some of the reported domains in ORF1 are RRM, CCHC, esterase, CTD and coiled coil domain [22]. The RRM domains identified in mammalian L1 ORF1 are non-canonical; they deviate significantly from the consensus sequence signatures, RNP1 and RNP2 of classical RRM domains [22]. Since traditional tools for domain prediction and ones which could find RRM domain in mammalian LINE1 ORF1were unable to find RRM domain in EhLINE1 ORF1, we used tools for RNA binding residue prediction (BindN and PPrint), which detected a long RNA-binding stretch at the N-terminal and some small stretches throughout the EhLINE1 ORF1. Both the tools predicted almost the same stretch, which lends confidence to the prediction. The only easily detectable domain in EhLINE1 ORF1p was a coiled coil domain towards the C-terminus. Typical coiled coil heptads [32] could be identified with high probability. The Multicoil tool also predicted a very high probability of dimer formation, and lower probability of trimer formation in this protein. However, this tool also gave the same prediction for human L1 ORF1p, which has been shown to form trimers [35]. The relative location of the RNA-binding and coiled coil domains in ORF1p of different LINE elements was not conserved [22], being in opposite parts of the molecule in EhLINE1 ORF1p compared with mammalian L1 ORF1p. Location of basic region in every element was also different, being at the N-terminus in some cases and C-terminus or middle in others (Fig. 1). This probably reflects the evolutionary course of different lineages of LINEs where domains of similar functions may have been assembled in different combinations.

Although ORF1p from a variety of LINEs lacks highly conserved domains, some properties of the protein, especially nucleic acid binding and chaperone activity seem to be conserved, as these may be required for RNA packaging during retrotransposition [13]. ORF1p is proposed to facilitate strand exchange and stabilize nick priming by annealing the poly (A) tail of human L1 RNA with T-rich region at the cleaved target site before reverse transcription [12,13,36]. We tested the nucleic acid binding ability of EhLINE1 ORF1p using a 65 nt RNA fragment from EhSINE1 3′ end, since it is similar in sequence to 3′ end of EhLINE1. This common 3′ end region may be recognized by EhLINE1 ORF1p and/or ORF2p to form RNP with LINE/SINE RNA for retrotransposition. We also tested the ability of this protein to bind ss and dsDNA corresponding to the same 65 nt sequence, and a 80 nt unrelated sequence from rRNA. While EhLINE1 ORF1p could efficiently bind RNA as measured by EMSA, it did not show appreciable binding to DNA. The K_d calculated for rRNA and SINE1-RNA was comparable. The K_d values are in agreement with those reported for other systems [11]. Although ORF1p had comparable affinity with unrelated RNA *in vitro,* its known function *in vivo*, namely that of forming RNP with LINE/SINE RNA to enable retrotransposition, would require specific recognition to exclude other cellular RNAs from being targets of retrotransposition. It is not known whether ORF2p, or some other cellular protein may impart this specificity. Specificity could also be imparted by specific tertiary structure of RNA *in vivo*. It may also be possible that ORF1p and LINE RNAs may be sequestered during translation itself, leading to the cis-preference observed for ORF1p for its own RNA [10].

Sequence analysis of EhLINE1 ORF1p showed that the putative RNA binding residues were located at amino acids 19–90, 117–125 and 268–280, while the coiled coil region was from 146 to 428 aa. Sub-fragments N-ter 23 (1–191 aa) and C-ter 37 (192–498 aa) did not show appreciable nucleic acid binding ability but N-ter 34 (1–277 aa) could bind efficiently to RNA. Although, the major stretch of RNA binding amino acids was included in N-ter 23, a smaller RNA binding region (268–280 aa) was present only in N-ter 34. It is possible that this extra stretch in N-ter 34 was required for proper folding of the polypeptide for nucleic acid binding. Another important difference between N-ter 23 and N-ter 34 is the presence in the latter of sufficient amino acid residues of the coiled coil domain to permit protein-protein interaction (Fig. 2), which was absent in the former. It has earlier been shown for mammalian L1 ORF1p that polymerization of the protein is needed for efficient RNA binding *in vitro* [10]. This could explain the inability of N-ter 23 to bind RNA in spite of having a substantial stretch of potential RNA-binding residues. The presence of nucleic acid binding property makes it likely that EhLINE1 ORF1p could perform retrotransposition when EhLINE1 ORF2p is supplied. This has, indeed been demonstrated *in vivo* in *E, histolytica* cells expressing both the polypeptides [19].

In spite of little sequence similarity the functional properties of EhLINE1 ORF1p resemble those of the ORF1p of other systems. This seems to be a common feature of ORF1p from other LINE elements as well. The ORF1p encoded by the zebrafish element ZfL2-1, which belongs to the esterase-type has no known RNA-binding domain; yet it possesses all the canonical activities associated with known ORF1ps, including self-interaction, nucleic acid binding and chaperone [14,37]. The ORF1p encoded by the *Drosophila melanogaster* LINE (I factor) contains a zinc-finger motif (CCHC) similar to the zinc fingers present in the basic region of retroviral gag polyprotein. It also has nucleic acid-binding and chaperone activity [25]. Thus it seems that the ORF1p of non-LTR retrotransposons has evolved a conserved function in spite of little sequence conservation, and our data with EhLINE1 also lend credence to this observation.

Thus far all the ORF1ps studied were from later-evolving non-LTR elements, while EhLINE1 belongs to the ancient R2 group that encode a REL-endonuclease domain. This domain was replaced in subsequent lineages with an APE endonuclease acquired from the DNA repair machinery of the host cell [4]. The acquisition of APE domain is believed to have coincided with the appearance of a second ORF (ORF1) 5′-of the major RT-encoding ORF (ORF2). In this respect EhLINE1 seems to be exceptional in that it has acquired the

ORF1 in a lineage that still contains the REL-endonuclease. Phylogeny based on ribosomal DNA sequences places *E. histolytica* close to the amoeba-flagellate *N. gruberi* [38]. Our analysis suggests that the ORF1p sequences of LINEs from these two organisms show conserved patterns (Fig. 1), which is significant as the two belong to different groups, and *N. gruberi* ORF2 encodes APE endonuclease. Conversely, the *Trypanosoma cruzi* LINE, L1Tc which encodes APE endonuclease, lacks ORF1 and has nucleic acid binding and chaperone activity at the C-terminal of its single ORF [33,39].

Phylogenetic analysis of non-LTR elements suggests that they have predominantly undergone vertical transmission as the oldest lineages of eukaryotes also harbor the oldest lineages of non-LTR retrotransposons [9]. However, these phylogenetic analyses are invariably based on the RT domain which is the best conserved. The 5′-part of the elements is highly divergent in sequence and has been left out in phylogenetic comparisons. Our data suggest that the vertical mode of inheritance of non-LTR elements holds for the 3′-part of the molecule but the 5′-part has evolved independently and has acquired divergent sequences while maintaining functional conservation necessary for successful retrotransposition.

Acknowledgements

A.K.G. and J.K. are recipient of Senior Research Fellowships from ICMR, M.A. is a recipient of Senior Research Fellowship from CSIR. V.P.Y. received research grant from DST. S.B. received research grant from DBT.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molbiopara.2016.11.004.

References

[1] P.R. Cook, C.E. Jones, A.V. Furano, Phosphorylation of ORF1p is required for L1 retrotransposition, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 4298–4303, http://dx.doi.org/10.1073/pnas.1416869112.
[2] D.C. Hancks, H.H. Kazazian, Active human retrotransposons: variation and disease, Curr. Opin. Genet. Dev. 22 (2012) 191–203, http://dx.doi.org/10.1016/j.gde.2012.02.006.
[3] B. Brouha, J. Schustak, R.M. Badge, S. Lutz-Prigge, A.H. Farley, J.V. Moran, H.H. Kazazian, Hot L1 s account for the bulk of retrotransposition in the human population, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 5280–5285, http://dx.doi.org/10.1073/pnas.0831042100.
[4] H.S. Malik, W.D. Burke, T.H. Eickbush, The age and evolution of non-LTR retrotransposable elements, Mol. Biol. Evol. 16 (1999) 793–805, mveb/16/604.793/805.
[5] T.H. Eickbush, H. Malik, Origins and evolution of retrotransposons, Mobile DNA ii. 2 (2002) 1111–1144.
[6] V.V. Kapitonov, S. Tempel, J. Jurka, Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences, Gene 448 (2009) 207–213, http://dx.doi.org/10.1016/j.gene.2009.07.019.
[7] J.N. Volff, C. Korting, A. Froschauer, K. Sweeney, M. Schartl, Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates, J. Mol. Evol. 52 (2001) 351–360.
[8] K.K. Kojima, K.I. Kuma, H. Toh, H. Fujiwara, Identification of rDNA-specific non-LTR retrotransposons in cnidaria, Mol. Biol. Evol. 23 (2006) 1984–1993, http://dx.doi.org/10.1093/molbev/msl067.
[9] W.D. Burke, H.S. Malik, S.M. Rich, T.H. Eickbush, Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia, Mol. Biol. Evol. 19 (2002) 619–630, http://dx.doi.org/10.1093/oxfordjournals.molbev.a004121.
[10] K.E. Callahan, A.B. Hickman, C.E. Jones, R. Ghirlando, A.V. Furano, Polymerization and nucleic acid-binding properties of human L1 ORF1 protein, Nucleic Acids Res. 40 (2012) 813–827, http://dx.doi.org/10.1093/nar/gkr728.
[11] V.O. Kolosha, S.L. Martin, High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1), J. Biol. Chem. 278 (2003) 8112–8117, http://dx.doi.org/10.1074/jbc.M210487200.
[12] S.L. Martin, The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition, J. Biomed. Biotechnol. (2006) (2006) 1–6, http://dx.doi.org/10.1155/JBB/2006/45621.
[13] S.L. Martin, F.D. Bushman, Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1, Mol. Cell. Biol. 21 (2001) 467–475, http://dx.doi.org/10.1128/MCB.21.2.467.
[14] M. Nakamura, N. Okada, M. Kajikawa, Self-interaction, nucleic acid binding, and nucleic acid chaperone activities are unexpectedly retained in the unique ORF1p of zebrafish LINE, Mol. Cell. Biol. 32 (2012) 458–6469, http://dx.doi.org/10.1128/MCB.06162-11.
[15] A.A. Bakre, K. Rawal, R. Ramaswamy, A. Bhattacharya, S. Bhattacharya, The LINEs and SINEs of Entamoeba histolytica: comparative analysis and genomic distribution, Exp. Parasitol. 110 (2005) 207–213, http://dx.doi.org/10.1016/j.exppara.2005.02.009.
[16] H. Lorenzi, M. Thiagarajan, B. Haas, J. Wortman, N. Hall, E. Caler, Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species, BMC Genomics 9 (2008) 595, http://dx.doi.org/10.1186/1471-2164-9-595.
[17] K. Van Dellen, J. Field, Z. Wang, B. Loftus, J. Samuelson, LINEs and SINE-like elements of the protist Entamoeba histolytica, Gene 297 (2002) 229–239, http://dx.doi.org/10.1016/S0378-1119(02)00917-4.
[18] R. Sharma, A. Bagchi, A. Bhattacharya, S. Bhattacharya, Characterization of a retrotransposon-like element from Entamoeba histolytica, Mol. Biochem. Parasitol. 116 (2001) 45–53, http://dx.doi.org/10.1016/S0166-6851(01)00300-0.
[19] V.P. Yadav, P.K. Mandal, A. Bhattacharya, S. Bhattacharya, Recombinant SINEs are formed at high frequency during induced retrotransposition in vivo, Nat. Commun. 3 (2012) 854, http://dx.doi.org/10.1038/ncomms1855.
[20] P.K. Mandal, A. Bagchi, A. Bhattacharya, S. Bhattacharya, An Entamoeba histolytica line/sine pair inserts at common target sites cleaved by the restriction enzyme-like line-encoded endonuclease, Eukaryot. Cell. 3 (2004) 170–179, http://dx.doi.org/10.1128/EC.3.1.170-179.2004.
[21] V.P. Yadav, P.K. Mandal, D.N. Rao, S. Bhattacharya, Characterization of the restriction enzyme-like endonuclease encoded by the Entamoeba histolytica non-long terminal repeat retrotransposon EhLINE1, FEBS J. 276 (23) (2009) 7070–7082.
[22] E. Khazina, O. Weichenrieder, Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 731–736, http://dx.doi.org/10.1073/pnas.0809964106.
[23] S.L. Martin, Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells, Mol. Cell. Biol. 11 (9) (1991) 4804–4807.
[24] V.O. Kolosha, S.L. Martin, In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 10155–10160, http://dx.doi.org/10.1073/pnas.94.19.10155.
[25] A. Dawson, A LINE-like transposable element in Drosophila, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids, EMBO J. 16 (14) (1997) 4448–4455.
[26] Y. Xiong, W.D. Burke, J.L. Jakubczak, T.H. Eickbush, Ribosomal DNA insertion elements R1Bm and R2Bm can transpose in a sequence specific manner to locations outside the 28S genes, Nucleic Acids Res. 16 (22) (1988) 10561–10573.
[27] S.M. Christensen, A. Bibillo, T.H. Eickbush, Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction, Nucleic Acids Res. 33 (2005) 6461–6468, http://dx.doi.org/10.1093/nar/gki957.
[28] V.K. Jamburuthugoda, T.H. Eickbush, Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase, Nucleic Acids Res. 42 (2014) 8405–8415, http://dx.doi.org/10.1093/nar/gku514.
[29] I.R. Arkhipova, H.G. Morrison, Three retrotransposon families in the genome of Giardia lamblia: two telomeric, one dead, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 14497–14502, http://dx.doi.org/10.1073/pnas.231494798.
[30] D.F. Estrada, D.M. Boudreaux, D. Zhong, S.C. St Jeor, R.N. De Guzman, The hantavirus glycoprotein G1 tail contains dual CCHC-type classical zinc fingers, J. Biol. Chem. 284 (2009) 8654–8660, http://dx.doi.org/10.1074/jbc.M808081200.
[31] J. Söding, A. Biegert, A.N. Lupas, The HHpred interactive server for protein homology detection and structure prediction, Nucleic Acids Res. 33 (2005) 244–248, http://dx.doi.org/10.1093/nar/gki408.
[32] D. Krylov, C.R. Vinson, Leucine zipper, Encycl. Life Sci. (2001) 1–7, http://dx.doi.org/10.1038/npg.els.0003001.
[33] S.R. Heras, M.C. López, J.L. García-Pérez, S.L. Martin, M.C. Thomas, The L1Tc C-terminal domain from Trypanosoma cruzi non-long terminal repeat retrotransposon codes for a protein that bears two C2H2 zinc finger motifs and is endowed with nucleic acid chaperone activity, Mol. Cell. Biol. 25 (2005) 9209–9220, http://dx.doi.org/10.1128/MCB.25.21.9209-9220.2005.
[34] S. Bhattacharya, A.A. Bakre, A. Bhattacharya, Mobile genetic elements in protozoan parasites, J. Genet. 81 (2) (2002) 73–86.
[35] E. Khazina, V. Truffault, R. Büttner, S. Schmidt, M. Coles, O. Weichenrieder, Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition, Nat. Struct. Mol. Biol. 18 (2011) 1006–1014, http://dx.doi.org/10.1038/nsmb.2097.
[36] S.L. Martin, M. Cruceanu, D. Branciforte, P. Wai-lun Li, S. Kwok, R. Hodges, et al., LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein, J. Mol. Biol. 348 (3) (2005) 549–561.
[37] S.L. Martin, J. Li, J.a. Weisz, Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1, J. Mol. Biol. 304 (2000) 11–20, http://dx.doi.org/10.1006/jmbi.2000.4182.

[38] M.L. Sogin, Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs, Integr. Comp. Biol. 29 (1989) 487–499, http://dx.doi.org/10.1093/icb/29.2.487.
[39] S.R. Heras, M.C. Thomas, F. Macias, M.E. Patarroyo, C. Alonso, M.C. Lopez, Nucleic-acid-binding properties of the C2-L1Tc nucleic acid chaperone encoded by L1Tc retrotransposon, Biochem. J 424 (2009) 479–490, http://dx.doi.org/10.1042/BJ20090766.
[40] A.K. Gupta, S.K. Panigrahi, A. Bhattacharya, S. Bhattacharya, Self-circularizing5′-ETS RNAs accumulate along with unprocessed pre ribosomal RNAs ingrowth-stressed Entamoeba histolytica, Sci. Rep. 2 (2012) 303.
[41] L. Wang, S. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, Nucleic Acids Res. 34 (2006) W243–W248 (Web Server).
[42] M. Kumar, M.M. Gromiha, G. Raghava, Prediction of RNA binding sites in a protein using SVM and PSSM profile, Proteins: Struct. Funct. Bioinf. 71 (1) (2008) 189–194.
[43] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, et al., Protein identification and analysis tools on the ExPASy server, in: The Proteomics Protocols Handbook, Springer, 2005, pp. 571–607.
[44] J. Schultz, F. Milpetz, P. Bork, C. Ponting, SMART, a simple modular architecture research tool: identification of signaling domains, Proc. Natl. Acad. Sci. U. S. A. 95 (11) (1998) 5857–5864.
[45] E. Wolf, P. Kim, B. Berger, MultiCoil. A program for predicting two-and three-stranded coiled coils, Protein Sci. 6 (6) (1997) 1179–1189.
[46] A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences, Science 252 (5009) (1991) 1162–1164.
[47] W. Burke, C. Calalang, T.H. Eickbush, The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme, Mol. Cell. Biol. 7 (6) (1987) 2221–2230.