CHAPTER 4

REPETITIVE DNA: CLONING AND CHARACTERIZATION OF A RETROTRANSPOSON-LIKE ELEMENT OF ENTAMOEBA
4.1 Introduction

4.1.1 Repetitive DNA elements
Repetitive DNA sequences are interspersed throughout the genomes of mammals, higher eukaryotes and also prokaryotes. They are referred to as "selfish DNA, because their functional implications remain unknown and they are maintained solely by their ability to replicate within the genome (Doolittle et al., 1980). There are several sequence classes amongst eukaryotic DNA. The constituent members can be broadly classified as (a) Unique (approximately one copy per haploid genome) e.g. protein coding genes, although more than one is possible (b) moderately repetitive (~$10^3$-$10^5$ copies per haploid genome). They are interpersed with unique sequences. Some of these dispersed repeats are mobile and cause recognizable mutations at their sites of insertion into chromosomal DNA. They are transcribed into discretely sized RNA molecules that have modified 3' and 5' ends, characteristic of cellular and viral mRNA. The short dispersed repeats can be categorized into different sequence families with different numbers of individual family members. Some families contain as many as $5 \times 10^6$ members, while others contain relatively few (Jelinek et al., 1982). (c) Highly repetitive DNA (approximately $10^6$ copies per haploid genome). The highly repetitive DNA consists of clustered repetitions of relatively short sequence units that are not interspersed with other sequence types. They are generally thought to be structural components of chromosomes, residing mainly at centromeric and telomeric positions, but their functions are unknown. They are organized as tandem repeats and referred to as satellite DNA and present as clusters in the chromosomes. The repeat size may vary from 5-10 bp to 100 bp with cluster sizes of upto 100 megabases. Microsatellite/simple sequences have 2-5 bp short repeats, example in vertebrates, insect, human and plant genomes. Mini satellites have 15 bp repeats with array size ranging from 0.5-30 kb. (Charlesworth et al., 1994).

4.1.2 Types of transposable elements
Movement of genetic information from one locus to another is known as transposition, and in principle, the information could be carried from the parental
locus to the target locus either by an RNA intermediate (retrotransposition) or via excision and reintegration of DNA itself. Although DNA mediated transposition has been documented in both eukaryotes and prokaryotes, RNA-mediated transposition appears to be restricted to eukaryotes. The DNA transposable elements are characterized by short inverted terminal repeats, and move by excision and reintegration (Fig. 6, Types of transposable elements). On the basis of similarity between the transposases, most eukaryotic transposons fall into two classes: the Ac/hobo class, characterized by an 8 bp insertion site duplication, and the Tc1/mariner class, with a TA dimer duplication. The presence of mariner-like elements in the mammalian genome has been published by several groups (Auge-Gouillou et al., 1995, Morgan et al., 1995, Oosumi et al., 1995, Smit et al., 1996). The human genome contains remains of DNA transposons (Engels et al., 1990). Retrotransposons are further divided into two super families based on common structural features: 1) the viral super family that encodes for reverse transcriptase (RT), this family includes retroviruses, long terminal repeat (LTR) retrotransposons, and long interspersed nucleotide elements (LINEs) also known as non-LTR retrotransposons. 2) the non-viral super family that does not encode for RT, which includes short interspersed nucleotide elements (SINEs) and processed retropseudogenes. SINEs are typically less than 500 bp in length and are characterized by an internal (mobile) Polymerase III promoter that ensures a fair chance for transcriptional activity of new copies. This promoter is present in a tRNA-derived region in all SINEs (Shedlock et al., 2000, Okada et al., 1991, Yoshioka et al., 1993, Kachroo et al., 1995), except in the primate Alu and rodent B1 SINEs, which have a common origin in a 7SL RNA derived Alu element (Quentin et al., 1994). Members have well defined 3' and 5' ends and there is sequence variation among members. An oligo A is present at the 3' end. SINEs usually make target site duplication of 7-21 bp upon insertion. The propensity of SINEs to insert into 3' terminal simple sequences (oligo A) of other SINEs can generate mobile composite elements (Weiner, 2000, Smit, 1996, Singer, 1982). The 6-8 kb LINEs are present in upto 500,000 copies in the human genome (Smit, 1999). The LINE1 encodes a RT and other proteins necessary for retrotransposition. An internal polymerase II promoter, present in at least some LINE-like elements, assures expression that is
Figure: 6. Types of transposable elements. The various types of transposable elements are shown in the figure.
relatively position – independent. SINES are distinguished in part from LINEs by their large copy number, relatively short length, and inability to encode for enzymes, such as RT, that are essential for their own amplification. The precise biochemical mechanism of SINE amplification is not yet completely known. The basic aspects are well characterized, however, and there is increasing evidence that SINES may acquire the necessary enzymes for retroposition from corresponding LINEs, which do code for RT (Eickbush et al., 1992, Okada et al., 1997, Oshima et al., 1996) and are thus capable of self-replication (Fig. 7, Sine amplification).

Wide spread distribution of both types of retrotransposons has been found in eukaryotes like fish (Flavell et al., 1992, Britten et al., 1995, Tristem et al., 1995), Ty1/copia like elements of amphibians, and reptiles (Flavell et al., 1995).

4.1.3 Mechanisms of transposition

4.1.3.1 Transposition by DNA breakage and rejoining

Transposition reactions occur within elaborate protein-nucleic acid complexes containing the DNA substrates and oligomers of transposase, and some times accessory proteins as well; assembly of these structures to form an active complex involves multiple protein-protein interactions and protein-DNA interactions. The functional assembly of these complexes is a key regulatory step in transposition. The requirement for proper assembly of such a complex prior to activation of catalytic activity assures that the appropriate substrates i.e. both ends of the mobile element and sometimes the target as well, are present and also promotes coordination of the reaction steps (Bainton et al., 1993, Mizuuchi et al., 1992, Chaconas et al., 1996, Craigie et al., 1996). The recombination complexes of Mu have been analyzed in some detail. Most notably, the active transposase is a multimer in which the active sites for DNA breakage and joining are formed by collaboration between domains from different proteins (Chaconas et al., 1996, Craigie et al., 1996). Distinctive protein-DNA complexes also underlie IS1 transposition (Kleckner et al., 1996). The nucleoprotein particles of retroviruses and LTR-containing retrotransposons are particularly elaborate, containing a variety of other proteins in addition to integrase and RT (Boeke et al., 1997).
Figure: 7. General Model for a possible tRNA-derived SINE amplification process. Corresponding LINE and SINE components are color-coded and share a common region (green) due to a recombination process. Reverse transcriptase (yellow) is generated by a LINE and the corresponding SINE transcript can be recognized by its LINE-derived tail region (bold black 3' end). The SINE transcript is then reverse transcribed into cDNA and integrates into the host genome (red site) by the target DNA-primed mechanism adopted by LINEs.
For most transposable elements, there is a highly sequence-specific interaction of the transposase with the recombination sequences at the ends of transposon, which are usually arranged as inverted repeats and are essential for substrate function. Once transposase is positioned at the transposon termini and assembled into an active complex, it executes the DNA breakage reactions that cut the tips of the transposon away from flanking DNA at the donor site, and then joins these exposed ends to the target DNA. In all systems that have been analyzed in biochemical detail, the critical chemical steps in transposition are DNA breakage reactions, which disconnect the 3'-ends of the transposon from the flanking donor DNA, and strand transfer reactions, which join these exposed 3'-ends to the target (Mizuuchi et al., 1992). The strand transfer step of transposition does not proceed via cleavage of the target DNA followed by ligation of the transposon ends into this broken target. In all cases, the joining of the transposon ends to the target DNA likely occurs via one-step transesterification in which an exposed 3'-OH transposon end acts as a nucleophile that attacks the target DNA, creating a new transposon-target junction with the concomitant production of an exposed 3'-OH in the target DNA (Mizuuchi et al., 1992, Mizuuchi et al., 1991, Engelman et al., 1991). The duplication of target site sequences characteristic of transposition followed from the repair of the small gaps that result from the joining of the two transposon ends at staggered positions on the target DNA (Benjamin et al., 1992, Benjamin et al., 1989, Bainton et al., 1991).

Many elements, including Tn10 (Benjamin et al. 1992), Tn7 (Hagemann et al. 1993), P elements (Kaufman et al., 1992), Tc1 (Vos et al., 1996) and other members of the widespread Tc1/mariner superfamily, and many plant elements as well, move by an excision-integration or cut-and-paste pathway in which the element is excised from the donor site by double-strand breaks that disconnect both the 3' and 5'-tips of the transposon from the donor backbone. A consequence of these double-strand breaks is the concomitant introduction of a gap into the donor DNA, reflecting transposon excision. How this donor gap is processed is also a part of the cellular response to transposition. Homologous recombination with a sister chromosome may repair the gap, a reaction seen with the Drosophila P element (Engels et al., 1992), with the nematode Tc1 element (Fischer et al. 1999), and with the bacterial Tn7 element (Hagemann et al., 1993). In some systems,
transposition occurs just after DNA replication, so an intact donor duplex always remains to repair the donor backbone gap resulting from element excision or to persist as an undamaged DNA (Bender et al., 1986). Rejoining of the flanks of gapped donor can also occur, although such reactions are most obvious in nonbacterial systems (Rommens et al., 1993). (Fig.8, DNA transposition)

4.1.3.2 Translocation of retroviruses and retroviral-like elements

The retroviral life cycle begins with the transcription of a full-length RNA that is then reverse transcribed, usually by an element encoded RT, to generate a double-stranded viral DNA (Boeke et al., 1997). This reverse transcription generally occurs within an elaborate nucleoprotein complex that, in the retroviral life cycle, becomes enveloped, is released from the cell, and infects a new recipient cell into whose genomic DNA integration will occur. With retroviral-like transposable elements, a nucleoprotein complex is formed that does not become enveloped and is not released from the cell; thus these elements are not infectious, and integration is intracellular. The element-encoded integrase usually initiates recombination by DNA cleavage events a few nucleotides from the 3'-ends of the viral DNA, exposing the actual 3'-OH ends of the transposable segment; these exposed 3'-OH transposon ends are then joined to the target DNA.

It is notable that in both the excision-integration and retroviral pathways, the key recombination intermediate is a double-stranded DNA form of the mobile element that is chemically disconnected from both the donor and target DNAs. In both cases, the exposed 3'-ends of this intermediate result from transposase directed cleavages; the free 5'-ends can result either from cleavage, as in the excision-integration pathway, or by reverse transcription, as in the retroviral pathway. These pathways differ, however, in their consequences at the donor site; the parental "proviral" segment that served as the transcriptional template for making the retroviral-like element RNA is not altered by the transposition process, whereas the excision integration pathway results in a gap at the donor site. Some bacterial elements, including bacteriophage Mu (Craigie et al., 1985) and elements of the widespread Tn3 (Roseman et al., 1995), remain attached to the donor DNA while inserting into a new target site. In these cases, a 3' nicking reaction occurs at each transposon
DNA transposition by excision and integration into a target site.

Transposase gene

GACTA

TARGET

ATGCA

Transposition

5'

3'

ATGCA

GACTA

ATCAG

ATGCA

Target site duplication

Figure 8. The mechanism of DNA transposition. DNA transposition by a 'cut and paste' mechanism by excision and integration into a target site.
end (rather than double strand breaks), so the 5'-ends of the element remain attached to the donor DNA. The exposed 3'-ends of the element then join to the target DNA. This transposition reaction is then followed by DNA replication to generate two copies of the element, one end of each copy being attached to the donor site and the other end to the target site, in a structure called a cointegrate. In some cases, the cointegrate is resolved by subsequent recombination between the element copies to generate two plasmids, one a target plasmid containing a simple insertion and the other a copy of the original donor plasmid (Craig, 1997). Thus, all these DNA transposition reactions are related by common 3'-end breakage, whereas the processing events at the 5'-ends may vary.

4.1.3.3 Transposition by DNA breakage and target priming for reverse transcription

Another large class of mobile elements, often called non-LTR or PolyA\(^+\) retrotransposons, move by reactions in which RNA plays a far more central role. These elements translocate by cleavage of the target DNA to produce a priming site for a RT that uses an RNA copy of the mobile element to generate in situ a DNA copy of the element linked to the new insertion site (Eickbush et al., 1992, Boeke et al., 1997). In some cases, the target site used is highly specific, insertion being observed only at a particular target site; this target site is specifically recognized by an element encoded endonuclease. In other cases, insertion is far less specific and is due to the action of a far less selective target nuclease. Non-LTR elements are widespread, having been found in many different organisms. Indeed, a considerable fraction of the genomes of many organisms are inactive copies of such elements for example, the human genome contains some 500,000 truncated copies and 3000-5000 full-length copies of the PolyA\(^+\) LINE (Smit, 1999). PolyA\(^+\) elements are characterized by the lack of LTR segments, i.e. they do not have special recombination sequences arranged in inverted orientation at their ends (Boeke et al., 1997). They do contain polyA sequences at their 3'-ends, and their 5'-ends are often truncated; it is not yet understood how the RT/endonuclease recombinase specifically recognizes the element RNA. Although PolyA\(^+\) elements generally do make target duplications, these duplications vary in length. Intact
copies of PolyA\textsuperscript{+} elements encode one or two proteins that contain RT and nuclease activity. Examples of PolyA\textsuperscript{+} elements include the R1 and R2 elements of insects, the I Factor of Drosophila, and the human L1 family. Other related elements include mobile group II introns and probably the specialized Drosophila TART and Het-A elements that act as telomeres.

4.1.3.3.1 Some PolyA\textsuperscript{+} elements prefer particular insertion sites
Site-Specific PolyA\textsuperscript{+} elements that insert into specific target sites are often found within arrays of repeated genes in host genomes. Repeated sequences may be a likely place of integration for mobile elements because the presence of many gene copies assures that gene function for the organism is not lost when one copy is occupied by a mobile element (Eickbush et al., 1995, Burke et al., 1995). The specific insertion sites are chosen by a target site-specific, element-encoded endonuclease. The R1 and R2 elements that have invaded rDNA genes are found in many insects (Jakubczak et al., 1991, Besansky et al., 1992) and also in nematodes (Burke et al., 1995); the R2 element of the silkworm B. mori (R2Bm) (Burke et al., 1987) has been studied extensively. Other target site-specific elements, such as CRE1, are found in tandem arrays of the spliced leader exons of Trypanosomes (Aksoy et al., 1987, Villanueva et al., 1991). R4, is a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes such as Ascaris lumbricoides (Burke et al., 1995). A retrotransposon of the non-LTR repeat class from the human blood fluke Schistosoma mansoni shows similarities to the Chicken-Repeat-1-like elements of vertebrates. (Drew et al., 1997).

Much about the mechanisms by which PolyA\textsuperscript{+} elements translocate has been revealed by the biochemical analysis of the mechanisms of R2Bm, an element that encodes a single open reading frame (ORF). Inspection of this ORF, and similar ORFs in many other non-LTR retrotransposons, revealed the signature amino acid sequences of an RT (Xiong et al., 1990). Demonstration of actual RT activity from the R2Bm ORF (Luan et al., 1993) and from the related Trypanosome CRE1 (Gabriel et al., 1991) and human L1 (Mathias et al., 1991) element ORFs was a key step in understanding PolyA\textsuperscript{+} element translocation. Also critical was the demonstration that the R2Bm ORF also encodes a specific endonuclease that
attacks the insertion site within 28S rDNA in vitro. This specifically broken target site can be used in vitro as a primer for a reverse transcription reaction using the R2Bm RNA as a template to generate R2Bm DNA linked to the target site (Luan et al., 1993). These studies provided biochemical evidence that R2Bm is directed to its target site through specific recognition of that target site by an element-encoded endonuclease, and that reverse transcription is used in the translocation of PolyA⁺ elements (Eickbush et al., 1992). (Fig. 9, Line retrotransposition)

4.1.3.3.2 Telomere-specific elements
At their termini, Drosophila chromosomes have the PolyA⁺ retrotransposons called TART (Levis et al., 1993) and Het-A (Biessmann et al., 1992), rather than conventional telomeres. In addition to being located at "standard" chromosomal ends (Levis et al., 1993), these elements have been observed to transpose to the ends of broken chromosomes (Biessmann et al., 1992, Sheen et al., 1994), a reaction that may be related to the "healing" of such ends. One of these elements, TART, contains an ORF that has amino acid sequence homology with RT; the Het-A element seems to be lacking an RT motif (Sheen et al., 1994). An interesting hypothesis is that these elements substitute for more conventional telomeres (Levis et al., 1993, Mason et al., 1995, Pardue et al., 1996). Other retrotransposable elements are also observed in telomeric regions, including the B. mori element TRAS1 (Okazaki et al., 1995) and the Yeast Y5 element (Zou et al., 1996, Zou et al., 1995).

4.1.3.3.3 Some PolyA⁺ elements insert at many different target sites
Line Elements: L1, I Factors, and more PolyA⁺ elements that can insert into many different target sites are widespread; particularly well-studied elements include the human L1 element (Dombroski et al., 1991), the I factor of Drosophila (Jensen et al., 1991), and the TAD element of Neurospora crassa (Kinsey et al. 1993). PolyA⁺ elements are also found in plants (Wright et al., 1996). Intact, full-length copies of these elements usually encode two ORFs. The function of the ORF1 gene product has not been defined, although it does appear to have RNA binding activity (Hohjoh et al., 1996). The second ORF has RT homology (Xiong et al., 1990). Recent
Figure 9. A possible mechanism for reverse transcription and integration of a new copy of L1Hs. (a) The ORF2 product, with reverse transcriptase and nuclease domains, binds to both the target DNA and full-length L1Hs RNA. The nuclease domain then cleaves one strand of the target at the site of integration (b) DNA synthesis is initiated using the 3’ hydroxyl of the broken strand of the target DNA as primer and a short region of the opposite strand as template. This generates the first copy of the target-site duplication (c) DNA synthesis continues using the 5’ end of L1Hs RNA as template. (d) When DNA synthesis reaches the 3’ end of L1Hs RNA, the second strand of the target DNA is cut. (e) The 3’ end of the first strand L1Hs DNA is joined to the 5’ end of the target DNA. (f) Synthesis of the second strand of L1Hs is initiated using the new 3’ hydroxyl of target DNA as primer. The short exposed strand of target DNA is using as template to generate the second copy of the target-site duplication. Synthesis then proceeds using the first strand of L1Hs DNA as template. (g) Synthesis of the new copy of L1Hs is complete and is flanked by the target-site duplication. An arrowhead indicates the 5’ end of each strand of nucleic acid.
analyses have also identified an endonuclease motif in ORF2 that is related to the apurinic/apyrimidinic repair endonuclease (Feng et al., 1996). More over, functional tests have demonstrated the presence of RT (Dombroski et al., 1994) and endonuclease (Feng et al., 1996) activities.

4.1.4 Retrotransposons in Protozoans
A non-LTR retrotransposon L1Tc has been characterized in great detail from *T. cruzi*. This 5kb transcript is present in a high copy number and found dispersed throughout the *T. cruzi* genome. The transcript at the 3'-end has a fragment of E12A (another highly repetitive DNA sequence of *T. cruzi*) (Requena et al., 1994) and a ribosomal mobile element (RIME) like sequence at the 5'-end. This 5kb transcript contains three non-overlapping ORFs in different reading frames. ORF 2 and 3 show homology with *pol* and *gag*-encoded proteins of non-LTR retrotransposons (Martin et al., 1995). The RT shows homology to RTs of other non-LTR retrotransposons. ORF1 showed significant homology with the Ape protein family (Demple et al., 1991). It is a protein with apurinic-apyrimidinic nuclease activity (Olivares et al., 1997). In *T. cruzi* another LTR retrotransposon SIRE is present in 1,500-3,000 copies per genome, depending upon the strain studied, and it is distributed in all chromosomes. SIRE is present in the 3'-end of several mRNAs always transcribed from the sense strand, contributing the poly A tail in 63% of the cases. In another study characterization of VIPER (Vestigial interposed retroelement), a 2326-bp-long unusual retroelement was done (Vasquez et al., 2000). VIPER's 5'-end is formed by the first 182 bp of SIRE, whereas its 3'-end is formed by the last 220 bp of the element. Both SIRE moieties are connected by a 1,924 bp long fragment that carries a unique ORF encoding a complete RT RNAse H gene where 15 C-terminal amino acids are derived from SIRE's region II.

In the Trypanosomatid *Crithidia fasciculata* CRE-1 a site specific retrotransposon has been well characterized (Gabriel et al., 1991). It is a non-LTR retrotransposon which inserts between tandemly repeated genes. It showed the presence of a RNA-directed DNA polymerase activity. In *Crithidia* another site specific, 9595 bp long element CRE2 inserts between tandemly arrayed miniexon genes. The element is
flanked by 29 bp target site duplications but lacks 3' poly d(A) tract characteristic of most non-LTR retrotransposons.

In *T. brucei gambiense*, a non-LTR retrotransposon similar to mammalian LINEs has been characterized (Aksoy et al., 1990). This retrotransposon, SLACS (splice leader associated conserved sequence) has been found to be associated with the spliced-leader (SL) RNA gene cluster. There are nine copies of SLACS and DNA sequence analysis of one shows hallmarks of a Line-1 like element. A 49 bp target DNA duplication is found at its insertion site and the 3'-end is preceded by a Poly(A) stretch. Two putative ORFs span 75% of the element. ORF1 shows the presence of -Cys-His motif associated with retroviral gag polypeptides while ORF2 shows homology with RT sequences. The 5'-end has a repeated segment of 185 bp that varies in copy number in different SLACS insertions. In *T. brucei* (Hasan et al., 1984) a ribosomal mobile element (RIME) has been characterized. The total length is 1028 bases. It generates 7 bp direct repeats at its target site. Each copy is a dimeric structure, one end of each monomer consisting of a stretch of 14A residues preceded by a putative trypanosome Poly A signal. Six bp of unknown origin are found in the dimer between the two copies of the element. The element is present mainly as a monomer, contains a 160 amino acid ORF and is extensively transcribed from both strands. The ORF orientation is opposite to that of rDNA transcription. In *T. brucei* another retroposon, Ingi, is found in several 100 copies per genome (Vassella et al., 1996). The ends of this element consist of the two halves (RIME A & RIME B) of another retroposon (Hasan et al., 1984). Transcripts of Ingi/RIME are of a heterogeneous size, most longer than the element itself. Another Trypanosome repeat sequence (TRS) has been reported from *T.brucei* (Murphy et al., 1987). The genome contains 400 copies of TRS. Majority of the copies are 5.2kb and flanked by different separate halves of the RIME element, although a variant copy contains only the central 1.45kb portion and lacks RIME. The TRS are bordered by direct repeats of 4 bp. Some TRS elements contain an ORF with homology to RT. The last third of the putative protein (1651 amino acid ORF) encoded by TRS may exhibit DNA binding properties which may have a possible gene expression. In *T. brucei*, expression of a retroposon-like sequence
have been detected upstream of the putative *T. brucei* variant surface glycoprotein gene expression site promoter. (Lodes et al., 1993).

### 4.1.5 Repetitive DNA elements characterized from *E. histolytica*

Several types of repetitive DNA have been reported from *E. histolytica* although no retrotransposon-like element has been reported so far. A repetitive DNA family from *E. histolytica* containing *S. cerevisiae* Autonomous Replicating Sequence (ARS) consensus sequence was reported by Lohia et al., 1990. Characterization of another repetitive DNA element from the pathogenic strain HK-9 has been reported (Michel et al., 1992). A repetitive clone B/C was isolated from a λgt10 library. This was isolated from non-rDNA chromosomal DNA and was found to be present in all the *E. histolytica* strains tested, as well as in other species of *Entamoeba* i.e. pathogenic and non-pathogenic strains, *E. moshkovskii* and *E. invadens*. No signal was detected in Northern blots indicating that it is not transcribed in the trophozoite stage. It contains two types of related direct repeats, CTTATTATA is tandemly repeated 12 times and the sequence CTTTATTATTAT is present 10 times. It also has trinucleotide repeats (TTA)n, which indicate that it may be part of a family of satellite DNA. A novel 0.55 kb transcribed repeat element has also been reported (Cruz-Reyes & Ackers, 1995). An unusual 0.55 kb DNA repeat element specific to *E. histolytica* called as interspersed element (IE) has been characterized. Hybridization of labeled IE sequences to *E. histolytica* DNA, indicate that the IE are reiterated about 500 times per *E. histolytica* genome. A 964 bp repetitive DNA containing 9 internal tandem repeats associated with linear chromosomal DNA has been reported by (Huang et al., 1997). This repeat has 5 ORFs and nucleotide sequence comparison showed 53.7% identity over 881 bp with the mitochondrial ori DNA region of *S. cerevisiae*. A 57.2% identity over 851 bp is shared with *Leishmania tarentolae* kinetoplast mitochondrial DNA. This homology may be due to high AT content of the sequence. Both *E. dispar* and *E. histolytica* have been found to contain this repeat but the number of copies may be much less in *E. dispar*. Another repeat element HMc was reported from our laboratory (Mittal et al., 1994). It is a 2.3 kb element found in several *Entamoeba* strains and was estimated to be present.
in 25-30 copies per haploid genome. This element encoded a transcript of about 1.35 kb. We have carried out further experiments to characterize this element.

4.1.6 DNA transposon-like elements in Protozoans

Following the sexual phase of its life cycle, the hypotrichous ciliate *Euplotes crassus* transforms a copy of its chromosomal micronucleus into a transcriptionally active macronucleus containing short, linear, gene-size DNA molecules (Tausta et al., 1991). Tens of thousands of DNA breakage and joining, or splicing, events occur during macronuclear development. The DNA removed by such events includes transposon-like elements, referred to as Tec1 elements, as well as segments of unique DNA sequence, termed internal eliminated sequences (IESs). Both types of elements are bounded by short direct repeats. The *Euplotes crassus* Tec1 and Tec2 elements have a putative transposase coding region (Jahn et al., 1993). Developmentally excised sequences in micronuclear DNA of *Paramecium* have also been reported (Steele et al., 1994). DNA processing occurs in ciliates at autogamy and conjugation when new macronuclei are formed from micronuclei and old macronuclei degrade. Processing of micronuclear DNA consists of removal of certain internal sequences, chromosomal fragmentation, addition of new telomeres, and amplification. Aside from a recent brief report, internal eliminated sequences have not been described in *Paramecium*. Nine internal eliminated sequences are found within and near the gene that codes for surface protein A in *Paramecium tetraurelia*. Of these nine, seven are located within the translated portion of the gene, and all include short inverted terminal repeats. The characteristic sequence, TA appears at the boundaries of all of the internal eliminated sequences. Developmentally programmed DNA deletion in *Tetrahymena thermophila* by a transposition-like reaction pathway has been well characterized (Saveliev et al., 1996). A molecular description of key intermediates in the deletion of two internal eliminated sequences (IES elements), the M and R regions, during macronuclear development in *Tetrahymena thermophila* has been described. Transposon-like elements in ciliated protozoa have a common "D35E" motif indicating the presence of a transposase superfamily (Doak et al., 1994). The transposon-like elements TBE1, Tec1, and Tec2 of hypotrichous ciliated protozoa
appear to encode a protein that belongs to the IS630-Tc1 family of transposases. The *Anabaena* IS9895 transposase also is placed in this family. Most family members transpose into the dinucleotide target 'AA'. Alignments including the additional members, and also mariner elements, show that transposases of this family share strongly conserved residues in a large C-terminal portion, including a fully conserved dipeptide, Asp-Glu (DE), and a block consisting of a fully conserved Asp and highly conserved Glu, separated by 34 or 35 residues (D35E). This D35E motif is likely to be homologous to the previously characterized D35E motif of the family of retroviral-retrotransposons, integrases and IS3-like transposases.

In *Oxytrichia fallax* a novel family of micronuclear elements termed telomere-bearing elements (TBEs) has been described (Herrick et al., 1985). All 1900 family members are eliminated during macro-nuclear development. These may be generated by transposition events. Sequence comparison of the termini and flanks of the element with the corresponding empty site indicate that elements cause 3 bp target duplications (AAT) upon insertion; the 3 bp are part of the 5 bp target sequence, AATGA. Moreover, both elements carry 77 or 78 bp inverted terminal repeats. The tip of each inverted terminal repeat is the 17 bp telomere-like sequence 5' C1A4C4A4C4. At least half of the elements have these 17 bp or an extremely similar sequence. One possible pathway for transposition into new micronuclear sites starts in the developing macronucleus with excision to create a free linear form to which telomeres are added, followed by a low frequency of movement to the micronucleus, and insertion into the germ-line micronuclear DNA.

### 4.1.7 Evolutionary dynamics of repetitive DNA in eukaryotes

Repetitive DNA in most cases seems to be maintained solely by their ability to replicate within the genome. (Selifish DNA hypothesis) (Doolittle et al., 1980, Orgel et al., 1980). Far from conferring benefits, their behaviour can sometimes result in fitness loss to the host (Charlesworth et al., 1986). Some human genetic diseases are known to be due to mutations caused by insertion of transposable elements (Wallace et al., 1991), chromosomal rearrangements induced by recombination between repeated sequences (Charlesworth et al., 1994), or amplification of micro satellite sequences (Kuhl et al., 1993). It has often been
proposed that repetitive sequences are functionally important for the host organism (Britten et al., 1994) or are maintained because their mutagenic activities contribute to the long term evolutionary potential of the population (Charlesworth et al., 1994).

4.1.7.1 Tandemly repeated non-coding DNA

Processes that effect tandemly repeated non-coding DNA

The forces affecting changes in the simple DNA repeat sequences (microsatellites) are best understood. Strand slippage is the major cause of observed length polymorphism of microsatellites within populations. This process results in length changes of a few bases at a time. Several processes as described in Fig. 10 A may lead to expansion or contraction of arrays of minisatellite repeats. New repeats on a given chromosome can originate either from the same chromosome or from its homologue. At present there is no evidence for unequal exchanges at minisatellite sequences. The frequencies with which changes in array size occur at micro and mini satellite loci are higher than normal mutation rates (Charlesworth et al., 1994).

The forces governing satellite DNA evolution are less well understood mainly because large size of their clusters preclude direct experimental analysis. Replication slippage involves few bases at a time hence may not be crucial for satellite DNA amplification. It may be important in initial formation as satellite DNA has smaller subunits. The presence of extra chromosomal circular satellite DNA suggests the possibility that long satellite repeats may primarily be amplified by extra-chromosomal rolling circle replication, followed by reinsertion into the genome.

The occurrence of unequal exchange has been inferred from sequence analysis of satellites containing higher order repeat units. It is however, not known whether unequal exchange takes place between sister chromatids or homologues. Modelling of the evolutionary dynamics of tandem arrays suggests that accumulation of very highly repeated sequences is expected to occur only in regions with very low recombination and weak selective constraints on array length. Recombination is suppressed near centromeres and telomeres, as is indeed the case in many eukaryotes as it is selectively advantageous. Higher order repeats are preferentially formed in regions of low recombination and weak selective constraints on total array size, provided that crossing-over between repeats depends on short stretches of
Figure: 10. Factors affecting the copy number of repetitive DNA. Factors affecting the numbers of copies of tandem repeats (panel A) and transposable elements (panel B) are shown.
sequence identity rather than long segments of overall homology e.g. satellite DNAs. Unequal exchange is a strong long-range ordering force which can keep tandem arrays homogeneous, even if, levels of crossing over are low relative to mutation (Charlesworth et al., 1994).

4.1.7.2 Transposable elements (TE)
Much of the moderately dispersed DNA of eukaryotes appears to consist of TEs. The spread and maintenance of most transposable elements in host populations is probably maintained by vertical (germ line) transmission. Fig. 10 B shows the factors effecting transposable elements. The majority of mutations caused by element insertion and element mediated chromosomal rearrangements must be deleterious, as in the case for other forms of mutations. However, it is also argued that transposable element activity may cause mutations that could have important developmental and evolutionary effects and that the beneficial consequences of this activity ensures their persistence. In some species, the transcription and transpositions of some transposable elements are correlated with specific developmental stages perhaps suggesting functional importance for the host organism. This will not however explain the spread and maintenance of other transposable elements in the face of selection against majority of the insertions (Charlesworth et al. 1994).

In D. melanogaster insitu hybridization using transposable element probes to polytene salivary gland chromosomes show that element frequencies are nearly always very low at chromosomal sites into which insertions can occur. Other studies on Drosophila genomes show that transposable elements almost never occupy coding regions. Because the TEs are capable of inserting into coding regions, selection must rapidly eliminate such insertions from natural populations. Contrastingly in mammals the SINE and LINE elements are evidently often fixed at high frequencies at individual sites (Charlesworth et al. 1994).

Drosophila population data show that existence of predominantly low population frequencies of transposable elements imply the existence of deterministic forces that oppose the spread of elements by transposition. The mean copy number of most elements seem to be roughly at equilibrium between an increase in copy number
under replicative transposition and a decrease in copy number brought about by the other forces. The survey data of *Drosophila* show that TEs are significantly over abundant in chromosomal regions in which the rate of meiotic recombinational exchange is reduced, as expected if ectopic exchange is reduced in parallel with regular meiotic exchange. The reason for the association of TEs within and around recombination-suppressing chromosomal inversions relative to uninverted homologous regions is yet to be discovered (Charlesworth et al., 1994).

### 4.1.7.3 Evolution of SINEs and LINEs

SINEs and LINEs have been described in an earlier section. Many thousand copies of these elements have been found in eukaryotic genomes. LINE retrotransposition has been studied by supplying RT activity in trans using an expression construct driven by a powerful viral promoter (Esnault et al., 2000). Most of the transpositions generated by this method were normal. This finding has an implication that our genomes are littered with LINEs, most in various states of irrevocable delay. So, why don't defective LINE elements interfere with retroposition of healthy LINEs? If the products of a healthy LINE can mobilize any LINE transcript in trans then LINEs with defective ORFs could propagate as parasites of healthy LINEs, much as SINE transcripts are thought to do, or perhaps like defective interfering particles do during viral infections. Mutant LINEs may accumulate at the expense of healthy LINEs and the LINE lineage may head for extinction.

Remarkably LINEs may be saved from extinction by purifying selection. The data of Esnault et al. indicate that the LINE retropositional machinery does not preferentially act on LINE mRNAs (as expected), but rather on the particular mRNA that encodes the RT. This cis effect implies that LINEs encoding functional products will be preferentially retroposed, so that retroposed LINEs are effectively selected for full protein function. Conversely this provides a key quality check against passive retroposition of non-functional LINE RNAs by active LINE· elements in trans. This has a problem as some of the data suggested otherwise, for example, elasmobranchs where the SINEs are retroposed by partner LINEs and the two have identical 3'-ends (The mechanism is explained in an earlier section). It remains to
be seen whether partner SINEs and LINEs actually exist in primates (Esnault et al., 2000).

The SINEs present in up to 10,000 copies in eukaryotic genomes have been used for diagnosing common ancestry among host taxa with extreme confidence due to the irreversible and independent nature of their insertions. As such, they represent a powerful new tool for systematic biology that can be strategically integrated with other conventional phylogenetic characters, most notably morphology and DNA sequences (Shedlock et al., 2000). The use of SINEs for tracing the evolutionary patterns has been made by Gilbert and Labuda, 2000. They have characterized SINEs of the core marsupial family in egg-laying (Monotremes), pouched (Marsupials), and placental mammals.

4.1.8 Aims and objectives.
The presence of a repetitive DNA family in *E. histolytica* widely distributed across the genome led to curiosity regarding its role in the genome organisation of the organism as well as its functional role if any. To characterize this repetitive DNA element of *E. histolytica* we carried out numerous experiments with the following objectives:

1. The understand the chromosomal organization of the repeat.
2. The determine the copy number of the element.
3. To characterize a number of members which would also include the complete repetitive unit.
4. To study the evolution of this repeat unit and its variation among different strains and species.
4.2 Results

4.2.1 Complete nucleotide sequencing of HMc
The repetitive DNA element HMc was originally identified and partially characterized in our laboratory (Mittal et al., 1994). Though the cloned fragment was about 2.3kb in length, nucleotide sequence of only 1.7kb was known. Recently it was noticed that nearly 500bp of HMc has a high degree of similarity with the upstream region of cpn60 gene of *E. histolytica* (Fig. 11). The nucleotide sequence of the complete fragment (2266 bp) was carried out and the sequence is shown in Appendix I. In order to decipher the function of this element a comparison was made with known sequences in the database using the algorithm BLAST (Altschul et al., 1997). A significant degree of similarity (p= 1e-04, score= 41) was observed with RT of many organisms. The BLASTX results are shown in Fig. 12 A. Incidentally the region shown to be similar to RT did not show any large open reading frame (Fig. 12B) suggesting that it may not be functional any more. Henceforth we will refer to HMc as EhRLE1 (*E. histolytica* retrotransposon-like element 1).

4.2.2 Copy number of EhRLE1
The copy number of EhRLE1 was determined essentially as described (section 2.13 –chapter 2). Briefly, *E. histolytica* strain HM-1:IMSS genomic DNA and EhRLE1 insert DNA were spotted onto a nylon hybridization membrane at a concentration ranging from 15-500 ng and 2-15 pg respectively. The membrane was then hybridized with radiolabelled EhRLE1 probe. Similar blots were prepared for *E. histolytica* calcium binding protein (Yadava et al., 1997), AP50 adapter protein (unpublished) and actin (Azam et al., 1996) and hybridized to respective insert DNA probes. Equal loadings of genomic DNA were ensured for all blots. The copy number of these proteins are known. The hybridization data is shown in Fig. 13. The copy number for calcium binding protein, AP50 adapter protein and actin were estimated to be 1, 1 and 5 per haploid genome respectively which are consistent with the known experimental values. For EhRLE1 the number of copies was determined to be 140 suggesting that this element is present in high copy number in the *E. histolytica* genome. The calculations for the copy number estimation are given in Appendix II.
Figure 11: Alignment of HMc/EhRLE1 with cpn60 gene upstream region. The EhRLE1 sequence shows homology to the reported sequence of the cpn60 gene upstream region.
BLASTX RESULTS OF HMc

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
<th>Score</th>
<th>E (bits)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pir</td>
<td>jS60004</td>
<td>hypothetical protein - common roundworm retrotr...</td>
<td>41</td>
<td>1e-04</td>
</tr>
<tr>
<td>gb</td>
<td>AAF99785.1</td>
<td>AC012463_2 (AC012463) T2E6.4 [Arabidopsis th...</td>
<td>41</td>
<td>0.065</td>
</tr>
<tr>
<td>pir</td>
<td>jT30935</td>
<td>reverse transcriptase - Atlantic horseshoe crab...</td>
<td>37</td>
<td>0.067</td>
</tr>
<tr>
<td>gb</td>
<td>AAF44126.1</td>
<td>AF139376_1 (AF139376) reverse transcriptase [...</td>
<td>40</td>
<td>0.084</td>
</tr>
<tr>
<td>pir</td>
<td>jB34087</td>
<td>hypothetical protein (L1H 3' region) - human</td>
<td>39</td>
<td>0.19</td>
</tr>
<tr>
<td>prf</td>
<td>j1207289A</td>
<td>reverse transcriptase related protein [Homo s...</td>
<td>39</td>
<td>0.19</td>
</tr>
<tr>
<td>gb</td>
<td>AAC51273.1</td>
<td>(U93570) putative p150 [Homo sapiens]</td>
<td>39</td>
<td>0.25</td>
</tr>
<tr>
<td>sp</td>
<td>Q03279</td>
<td>PO21_SCICO RETROVIRUS-RELATED POL POLYPROTEIN FRO...</td>
<td>39</td>
<td>0.25</td>
</tr>
<tr>
<td>gb</td>
<td>AAD15471.1</td>
<td>(AC006067) putative non-LTR retroelement rev...</td>
<td>39</td>
<td>0.25</td>
</tr>
</tbody>
</table>

A

Figure : 12. BLASTX result and ORF status of HMc. Panel A: The BLASTX results for HMc. The homology with RT is indicated with several organisms. Panel B: Shows the ORF finder results for HMc. No large ORF is picked in the search.
Figure: 13. Copy number determination of EhRLE1. *E. histolytica* genomic DNA and purified insert DNA were spotted in duplicates in rows 1, 2 and 3, 4 respectively in panels A, B, C and D. The amount of genomic DNA spotted in rows 1, 2 is 500 ng to 15 ng in columns 1 to 6 for all panels. The quantities are indicated above the panels. Columns 7 and 8 of rows 1 and 2 have been loaded with PBS and PTZ18R vector DNA respectively. The insert DNA has been spotted in duplicates in rows 3 and 4. The amount ranges from 2 ng to 15 pg in columns 1 to 8 for all panels. The quantities are indicated below the panels. Hybridization has been done with EhRLE1 insert DNA as probe in panel D. Copy numbers of Calcium Binding Protein (Panel A), actin (Panel B) and clathrin (Panel C) have also been estimated in a similar manner as controls.
4.2.3 EhRLE1 is present in all chromosomes of *E. histolytica*

The chromosomes of *E. histolytica* can be separated by Pulsed Field Gradient Electrophoresis (PFGE) (Bagchi et al., 1999). In order to assign chromosomal location of EhRLE1, PFGE separated *E. histolytica* DNA from strains HM-1:IMSS and HK-9 was hybridized to radiolabelled EhRLE1 DNA as probe and the results are shown in Fig. 14. For comparison ethidium bromide stained PFGE patterns are also shown. All the PFGE bands were found to hybridize with the probe suggesting that EhRLE1 may be present in all chromosomal bands. This was further confirmed by comparing a densitometric scan of the ethidium bromide stained gel with the autoradiogram. The two showed excellent correspondence with each other (data not shown). This rules out any discrepancy arising out of band compression.

4.2.4 The whole EhRLE1 is part of the repetitive element

It is likely that only a part of EhRLE1 may be the repeating unit rather than the whole fragment. In order to test this 3'-end and the 5'-end fragments of EhRLE1 were used as probes to hybridize to EcoR1-digested genomic DNA. Normally under this condition and using complete EhRLE1 as a probe a large number of bands are observed indicating repetitive nature of the fragment (Fig 15, panel I). Both the 5'-end, the 3'-end and the complete EhRLE1 probes gave similar hybridization patterns of multiple bands suggesting that the entire EhRLE element may be part of the repetitive DNA (Fig 15, panel II and III). Similar analysis was also carried out using different restriction enzymes, Nde1 and Taq1 and results were essentially similar. There were a few fragments, such as 1.7 kb and 1.4 kb that showed much higher intensity in Taq1 and Nde1 digested DNA respectively (Fig.16). These may be derived from the internal portion of the repetitive element.

4.2.5 Molecular cloning and characterization of EhRLE2 and EhRLE3

The data shown so far indicate that the repetitive element EhRLE1 is present in about 140 copies and is distributed throughout the genome. It is likely that the 2.26 kb fragment may not be a complete repetitive unit. In order to characterise a complete repetitive unit attempts were made to identify and clone larger fragments carrying EhRLE1. A genomic library of *E. histolytica* consisting of EcoR1 digested
Figure: 14. EhRLE1 is present in all chromosomes of *E. histolytica*. The chromosomes of *E. histolytica* strains HM-1:IMSS and HK-9 were separated by pulsed field gel electrophoresis (PFGE). The DNA was blotted and hybridized to EhRLE1 probe and all chromosomes visible on ethidium bromide staining also gave a signal with EhRLE1. The ethidium bromide pattern is shown for comparison.
Figure: 15. The entire EhRLE1 is part of the repeat unit. E. histolytica strain HM-1:IMSS genomic DNA restriction digested with EcoR1 has been hybridized to indicated probes: Panel I: Total EhRLE1 probe, Panel II: 5'end probe A from EhRLE1 Panel III: 3'end probe B from EhRLE1. The positions of the probes are indicated in the map.
Figure: 16. EhRLE1 probe gives signals from multiple bands upon digestion of *E. histolytica* strain HM-1:IMSS DNA with several other restriction enzymes. HM-1:IMSS DNA was digested with several enzymes and hybridized to EhRLE1 probe. Lane1: Taq1, Lane 2: Nde1 and Lane3: Hpa1 digested DNA. Signals were obtained from bands ranging from over 10 kb to less than 1 kb. The 1.7 and 1.3 kb bands gave strong signals in Taq1 and Nde1 digested DNA.
fragments was screened using EhRLE1 as a probe. After the primary screening of 
40,000 plaques, a number of plaques were purified through secondary and tertiary 
screenings (Fig 17). Finally 9 recombinants were identified for further characterization 
by determining the sizes of the inserts released by digestion with the restriction 
enzymes EcoR1 and Nde1 and Southern hybridization by EhRLE1 as a probe. Only 
two clones EhRLE2 and EhRLE3 larger than the original EhRLE1 were found.

4.2.6 Characterization of EhRLE2 and EhRLE3
The complete nucleotide sequence of EhRLE2 and EhRLE3 ( appendix I ) was 
determined after a set of nested deletions were generated using Exolll (see materials 
and methods). EhRLE2 is of 2747 bp in length (Fig. 18). The comparison of the 
sequence with EhRLE1 showed that the 5'-end of EhRLE2 has an additional 223 
nucleotides and the identity at the 3'-end ended at nucleotide position 1469 (Fig. 18). 
EhRLE3, on the other hand is of 2473 nucleotides and identity with EhRLE1 started 
from nucleotide position 216 and continued till the 3'-end. There were minor 
sequence variations among all the three sequences suggesting that the three 
fragments may be from different genomic locations.

4.2.7 Molecular cloning of complete EhRLE repeating units
It is clear that both EhRLE2 and EhRLE3 do not have the complete repeating unit. In 
order to clone genomic fragments containing the entire repeating unit a library was 
constructed by digesting genomic DNA with EcoR1 and screened with labelled 
EhRLE1 as probe (Fig 19). Since Southern hybridization has already shown (Fig 15) 
presence of large fragments which contain this repetitive element it is likely that 
these may have the entire repetitive element. After screening and purification, 3 
independent clones were obtained. These clones were analyzed by restriction 
digestion and Southern hybridization and the data is summarized in Fig. 20. It 
appears that EhRLE1 may lie at the central region of EhRLE4 and EhRLE5. 
Therefore the entire repeat element may be present in these fragments. The smaller 
4.0 kb fragment EhRLE6 was different from the other two and it was not sure if the 
entire repeat unit is present in this fragment or not. In order to characterize in detail 
the complete nucleotide sequence of these clones were determined ( Appendix I ).
Figure: 17. Screening of *E. histolytica* strain HM-1:IMSS genomic library with EhRLE1 probe. The primary and secondary rounds of screenings of the HM-1:IMSS genomic library with EhRLE1 probe are shown in panel A and Panel B respectively.
Figure: 18. Domain maps of EhRLE2 and EhRLE3. The domain maps of EhRLE2 and EhRLE3 indicate the nucleotide positions of the various regions of homology in these clones and also their positions with respect to EhRLE1.
Figure: 19. Shotgun cloning and screening of genomic library of EcoR1 digested *E. histolytica* strain HM-1:IMSS DNA with EhRLE1 probe. HM-1:IMSS genomic DNA was digested with EcoR1 and ligated to PBS KS+ vector. *E. coli* DH5α cells transformed with the recombinant DNA were screened with EhRLE1 insert as probe. The positive clones obtained are indicated in the figure.
Figure: 20. Restriction mapping and Southern hybridization of EhRLE4, EhRLE5 and EhRLE6 with EhRLE1 probe. The positive clones obtained from the shotgun library were restriction mapped with EcoR1 and NdeI enzymes. The restriction map is shown at the bottom of the figure. The 3 clones (EhRLE4, EhRLE5 and EhRLE6) were digested with EcoR1 in lanes 1, 3 and 5 respectively. The same clones were digested with EcoR1 and NdeI (in the same order) in lanes 2, 4 and 6 respectively. The DNA was hybridized to EhRLE1 probe. The positions of the bands giving signals with EhRLE1 probe can be seen on the restriction map. The fragments giving signal with EhRLE1 are located internally in these large clones.
4.2.8 Molecular characterization of different EhRLE elements

The EhRLEs were analysed by pairwise comparison with known sequences in the databases using the algorithm BLAST (Table 6) and by multiple alignment using the program CLUSTALW (Fig. 21). Detailed maps along with characteristics of different sequence elements present in each of the EhRLE are shown in Fig. 22. Of the six EhRLE clones analysed, EhRLE5 appears to contain the complete EhRLE unit. Our data shows that the complete EhRLE unit is 4087 bp long. It is flanked at both ends by a 22-mer exact repeat which consists of a 8-mer direct repeat (AA TT AA TT) and a 14-mer inverted repeat (AAGTATT AA TAGAA). This is followed by a 641 bp stretch that shares 85% homology with a sequence reported to occur 326 bp upstream of the superoxide dismutase (SOD) gene of E. histolytica (Fig. 23, Tannich et al., 1991). After this is a stretch of 570 bp which shows no match in the database, followed by a 1343 bp sequence with strong homology (score= 79, P value= 2e-21) with RT. 3'- of the RT region is a 248 bp region that shares 88% homology with a sequence reported to occur 645 bp upstream of the cpn60 gene of E. histolytica (Clark et al., 1995). This is followed by another stretch of 759 bp with no match in the database, and finally a 506 bp region with 88% homology to the IE element of E. histolytica (Cruz-Reyes et al., 1995) after which is, the 22-mer repeat at the 3'-end. The 5'-end of the unit is flanked by a gene coding for the cation transporting ATPase. The 3'-end is flanked by a stretch of sequence with no homology in the database (Fig. 21).

The description given above is of the complete EhRLE unit contained in the clone EhRLE5. The clone EhRLE4 also contains an almost complete EhRLE unit, except that the IE element and the 22-mer repeat at the 3'-end are missing. The other clones contain shorter parts of the complete unit. However, EhRLE6 has an SOD upstream region and a very short truncated RT region (Fig. 24). The SOD upstream region sequence is similar to that of the other EhRLEs but the 22 bp repeat is truncated. Only 11 bases of this repeat are present (Fig. 25). Due to the presence of an extremely truncated RT region the sequence of this clone was not used for further analysis. At the nucleotide sequence level there was a high degree of identity among all EhRLEs. But the nucleotide sequence of no two EhRLEs was found to be
<table>
<thead>
<tr>
<th>EhRLE Type</th>
<th>Unique Region</th>
<th>Blast Score</th>
<th>P value</th>
<th>SOD Region</th>
<th>Blast Score</th>
<th>P value</th>
<th>No Match</th>
<th>RT Conserved Region</th>
<th>Blast Score</th>
<th>P value</th>
<th>cpn60 Region</th>
<th>Blast Score</th>
<th>P value</th>
<th>No match</th>
<th>IE element</th>
<th>Unique region</th>
</tr>
</thead>
<tbody>
<tr>
<td>EhRLE1</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-248 249-1321 41 1e-04 1295-1832 297 1e-77 1833-2266</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EhRLE2</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-471 472-1469 47 0.001 —</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EhRLE3</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-464 465-1788 84 1e-6 1510-2008 529 1e-147 2009-2473</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EhRLE6</td>
<td>1-156</td>
<td>—</td>
<td>—</td>
<td>157-783 357 7e-96 784-1136</td>
<td>315 3e-83 — — 1593-1940 unlike EhRLE</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EdRLE</td>
<td>2719-3054</td>
<td>—</td>
<td>—</td>
<td>— — —</td>
<td>— — —</td>
<td>—</td>
<td>—</td>
<td>2718-2540 2539-1467 52 5e-87 1466-967 163 6e-37</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: 6. BLAST homology scores and P values of EhRLE elements
Multiple Alignment of EhRLE sequences

START OF EhRLE4

5' END

EhRLE4

---

EhRLE5

---

EhRLE4

---

EhRLE5

---

EhRLE4

---

EhRLE5

---

EhRLE4

---

END

...
START OF EhRLE REPEAT

END OF UNIQUE SEQUENCE

START OF SOD UPSTREAM REGION

8 bp DIRECT 14 bp INVER
<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EhRLE1</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE2</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE3</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE4</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE5</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EhRLE4</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE5</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE2</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE3</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EhRLE4</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE5</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE2</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE3</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EhRLE4</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE5</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE2</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
<tr>
<td><strong>EhRLE3</strong></td>
<td>ATTCTGAAAT</td>
<td>CATCCATAA</td>
<td>TAA-GTGTGA</td>
<td>TTGAGATTGT</td>
<td>GGAATTTCT-</td>
<td>CTTCAAT- GAGTTGAT</td>
<td>TTGAGAAAT</td>
</tr>
</tbody>
</table>
Figure: 21. Multiple Alignment of EhRLE elements using the CLUSTAL Program. The various regions of sequence homology are shown with colour coded lines running on top of the sequence. The sequence of the 14 bp inverted repeat is coloured purple and the 8 bp direct repeat is coloured blue. The TATA-like box is shown in maroon. The nucleotides of each element varying from the consensus sequence are coloured. The underlined nucleotide positions are those without a consensus. Colour codes for the various regions are given below:

- **UNIQUE SEQUENCE**
- **cnfl60 UPSTREAM REGION**
- **SOD UPSTREAM REGION**
- **NO MATCH**
- **IE ELEMENT**
- **RT REGION**
- **UNIQUE REGION**
Figure: 22. Domain maps of all EhRLEs. The figure shows the maps of all EhRLE clones aligned with each other with respect to EhRLE1. The colour codes for homologous regions are same for all maps. The unique regions are shown with a different colour. The BLAST scores for the various regions of homology of the different clones are indicated in Table 6.
Figure 23: Alignment of SOD gene upstream region with homologous region from EhRLE5. A region of 641 bp located at the 5' end of the EhRLE unit shows homology to the upstream region of reported SOD gene sequence.
Figure: 24. The domain map of EhRLE6. A very short EhRLE-like region is present here. The SOD upstream region and a short RT region are shown in the map. An unsequenced region (red) is indicated in the map. The regions in brown do not show any homology to database sequences. The unique coding regions are indicated in the map.
ALIGNMENT OF EhRLE6 WITH EhRLE5

147  EhRLE6  156
AAAGATCCGT  START OF SOD UPSTREAM REGION
::::
AAAGATAAAAAATTCTATAAAAAATAAACAGAAAGATGAGAGAATAAAATAAGGA
1104  EhRLE5  1173

THE 22 bp REPEAT IS TRUNCATED IN EhRLE6  EhRLE6  212
TATTAATAGAAATTAGCTTCTATATTCTTCTGTCGTCAGGGAGATATTACCAAATGAG
::::
ATTAAATAGATATTTGAACTACTTCGTTCTATATTCTTCTTGAGATGAGAGAAAATAATG
EhRLE5  1243

EhRLE6  278
TCTTCTTGTAGCTCTAGGAGTATGGTTGTTGTGATCTCCGCAATA TCTATGACATCTCATGGT
::::
CATCTTCTAGCAGCTCTAGGATGTGGTTGTGATAGAAGCAGCATTCCATGACATTAATGATGTT
EhRLE5  1313

EhRLE61  329
TTTTATGACAAGGTTTTTTAAATTTTTTTGGATGTCATTTTTTAT TCTTT
::::
TGTTTATTCAATGTTTTTTAAT TCTTTAATGAGGTTTCAATTTTTAATTTCT
EhRLE5  1369

EhRLE6  387
TAAATAATTTTAACTTCTTATTTTTATC TAAATCCATTGAGGATGATACTATTGGAA
::::
TAAATACTGATAC TCTTTTGTATTATGTAATTTTTAATTTAT
EhRLE5  1426

EhRLE6  457
TATTTGACATCTTTTTATTTTTTTATCTATGCTATAGCTATTCTTATTACTATTATTTATTTT
::::
TTATT TGGACATCTTTTTATTTTTTTATCTATGCTATTCTTTATTATTGATATT
EhRLE5  1495

EhRLE6  527
TCTTCAAAATCACAATTTAATTTTGGATCAGCCACAAATGTCATATCGGAAATTAGACGACTTTTGAT
::::
TCTTCAAAATCACATTTAATTTTTGGATCAGCCACAAATGTCATATCGGAATATTAGACGACTTTTGAT
EhRLE5  1564

EhRLE6  596
TCTTTTTTATCAATAGCTATCTTCTGTATATTTTTTGTGATTGCAA CACTTTTATCATTATTTGATT
::::
TCTTTTTTATCAATAGCTATCTTCTGTATATTTTTTGTGATTGCAA CACTTTTATCATTATTTGATT
EhRLE5  1632

EhRLE6  664
GTCCCATATCATTGTTAGCTTTTATATATATCATTATTTTGTGATTGCAA TGAACATTTGTTCATATTTG
::::
GTCCCATATCATTGTTAGCTTTTATATATCATTATTTTGTGATTGCAA TGAACATTTGTTCATATTTG
EhRLE5  1702
EhRLE6 734
CATGTGTATTTTCTTATTAATCAATATTATAATGCGCAATATTTTCTTATTTTCCATTTTTGATTTTTTA

END OF SOD UPSTREAM REGION AND START OF REGION OF NO MATCH
EhRLE5 1772

EhRLE6 803
TGACAAATACAAGATAGTAGAATGTGCTCTACTGTAGCTATATTGTCTTACATAGTGTGGAGTAGTCTA

EhRLE6 1842

ATCT TCATTTCATATTTGCTTTATACCCACAAATATTGCTCCATCTGCATTAGTTAA

EhRLE5 1912

ATCTTTCATTTTTCATTATTTTTGATTTTCTATACCCACAAAATATTGCTCCATCTTGCATTTTAGTTAA

EhRLE5 1982

TTGTAGGAATGCATTCTTTTTAATATTGATGAAGTTCCATGCTTTAATTGTAGTAATTTTATCTACTCCT

EhRLE5 2052

TTTGACTCGAACCATATTCTAAAATATCTACTAATTTGATGACACTTCCATTTTTTGATGTATTTGTCA

EhRLE5 2122

ATAATTTTTTTATTCCTACTACTACTTCTTTTTCTTTTTTTCTTTTCAATAAATTCATTTATTATTTCT

EhRLE5 2192

ALIGNMENT OF EhRLE6 WITH EhRLE4

START OF SOD UPSTREAM REGION
EhRLE6 158
TTCTTTTTTCC TTTTTTTAAGAAAACATATAAAAACAAAGATCCGT TA

EhRLE4 1387

11 bp OF THE 22 bp REPEAT ARE PRESENT IN EhRLE6
EhRLE6 221
TTAATAGAATTAGCTTCTATTATTCGCTGCAGGGAGATTTAATCCAAATGA

EhRLE4 1457
FIGURE: 25. Alignment of EhRLE6 sequence showing homology to SOD upstream region with EhRLE5 and EhRLE4 SOD upstream region sequence. The 22 bp repeat (bold and underlined) found at the 5' end of the SOD upstream region of EhRLE5 and EhRLE4 is truncated in EhRLE6. Only 11 bp of the repeat are present.
identical. The percentage variation ranged from 1.9% (EhRLE3) to 4% (EhRLE5) from the consensus sequence (described in a later section). The percentage variation is 1% to 3% in the RT region for the same clones suggesting lesser variation in the coding region as compared to the overall sequence variation. This suggests that these EhRLEs have originated from different locations of the *E. histolytica* genome.

4.2.9 Characterization of the ends of the EhRLE repeat unit

Nucleotide sequence alignment suggested that the 5'-end boundary of the repeating unit is upstream of the SOD-homology region. This was confirmed by Southern hybridization with probes from SOD-homology region and from the 5'-end unique region (Fig. 26). The data shows single hybridization band with 5'-end unique region of EhRLE4 and EhRLE5 (panel II and III) and multiple bands with SOD-homology region similar to that obtained with EhRLE1 probe (panel I). There were a few more bands at the <2 kb region obtained with SOD probe compared to the EhRLE1 probe. This may be due to presence of SOD-upstream region in locations other than EhRLE, or due to presence of an internal EcoR1 site in this region in some EhRLE elements across the genome. Similarly the 3'-end was also analysed by Southern hybridization. The probe from within the repeat unit gave multiple bands after hybridization whereas unique region probes from both EhRLE4 and EhRLE5 showed a single band (Fig. 27). The data suggest that the ends inferred by sequence alignment are indeed the true ends of the repeat units.

The data presented so far indicate that EhRLE is probably a transposable element, and due to the presence of an RT could be a retrotransposon.

4.2.10 Site of integration of EhRLE

The integration of retrotransposons into the genome may be either site specific or random (Craig, 1997). Both the junctions of EhRLE4 and EhRLE5 were analyzed for presence of any unusual sequence organization. A 22 mer inverted repeat containing a 8-mer direct repeat was found at both the ends of EhRLE5 (Fig.21). These may be real integration sites as the probability of finding 22-mer inverted repeats in any genome is quite low. In EhRLE4 the 22 mer sequence was present
Figure: 26. Southern hybridization of *E. histolytica* genomic DNA with 5’end repeat and unique region probes. *E. histolytica* genomic DNA digested with EcoRI was hybridized to: Panel I: 5’end repeat region (SOD upstream region) (probe A), Panel II: 5’end unique region probe from EhRLE4 (probe B), Panel III: 5’end unique region probe from EhRLE5 (probe C). The positions of the probes are indicated in the map.
Figure: 27. Southern hybridization of *E. histolytica* strain HM-1:IMSS DNA with 3' end repeat and unique region probes. *E. histolytica* genomic DNA digested with EcoR1 was hybridized to: Panel I: 3'end repeat region (probe A), Panel II: 3'end unique region from EhRLE4 (probe B) and Panel III: 3'end unique region from EhRLE5 (probe C). The positions of the probes are indicated in the map.
only at the 5'-end. It is likely that the sequence at the 3'-end may have been lost by a deletion event.

4.2.11 Reconstruction of RT sequence

It has been pointed out before (see Fig. 12B) that there is no large ORF in the RT homology region of EhRLE1 inspite of considerable similarity with RT sequences. Similar results were found for all elements. It is likely that the sequences in present day EhRLEs may have been derived from a functional RT by accumulation of mutations. A reconstruction of what may be the original RT sequence was attempted by making changes (insertions and substitutions) in the consensus sequence. A consensus sequence was generated after multiple alignment of EhRLE nucleotide sequences except EhRLE6 (Fig. 28) and by comparing the RT homology region of all the EhRLEs and using EhRLE3 as the backbone as it was found to be closest to RT sequence by BLASTX analysis. Normally majority nucleotide was used in a given position if there was conservation. At positions where no majority could be arrived at, the nucleotide present in EhRLE3 was kept in the consensus. The sequence was searched for an ORF. The different reading frames of the sequence that showed identity with the RT sequence were analysed. It became clear that there was a shift of one nucleotide. When a 'T' was inserted at position 1365 of the consensus sequence a continuous uninterrupted reading frame consisting of 464 amino acids (nucleotide positions 1575-361) spanning the RT coding region was obtained (Fig.29). The modified consensus sequence (with continuous long ORF) was analysed using BLASTX and the match with RT sequences improved considerably as reflected in lowering "p" value (1e-26 (EhRLE3) to 3e-30 (consensus)) and the increased score (84 (EhRLE3) to 134 (consensus)).

4.2.12 Phylogenetic analysis of RT from EhRLE

Deduced amino acid sequence of the putative RT region of EhRLE was aligned along with a number of RT sequences of non-LTR retrotransposons from different species (Fig. 30). As pointed out in the figure some of the conserved RT domains are also conserved in EhRLE suggesting that these may have been derived from RT. Seven conserved amino acid positions found in functional RT sequences were found
Figure: 28. Construction of a consensus EhRLE sequence. The EhRLE Sequences were aligned together and a consensus sequence was generated keeping EhRLE3 as the backbone. The majority nucleotide was chosen from positions showing nucleotide conservation. At positions without conservation the nucleotide in EhRLE3 was kept in the consensus sequence. The consensus sequence is shown in maroon and a single nucleotide insertion (T) at position 1365 to generate a continuous ORF is shown in blue. EhRLE6 was excluded from the alignment due to it’s short length. The nucleotides of each element deviating from the consensus are coloured and positions without consensus are underlined. Nucleotide positions ‘1’ and ‘2485’ of the consensus sequence correspond to positions ‘2170’ and ‘4735’ respectively of the multiple alignment shown in figure. 21 for the entire EhRLE sequence.
CONSTRUCTION OF A CONTINUOUS RT ORF

Length: 464 aa  Reading frame : -3

1755 atgtttgattctacaagaaaataggataggagaagaaagaccca
    MF DSNK KIRIRK KDP
1770 aaaaatgaataatcaccaaataacgatatcttttagaatatttg
    KNEIYPNNDITYW
1665 aaatgtttatagaaacaggtgtatattaaataaagaaaattgg
    KSLYETQVILKENW
1620 aaaaataaaaccaatcctcggttgaagatagatctataaatat
    KIKQITSWENERNY
1575 gatagttcataataacataaataaatgagataatattatg
    DSVITINEINALSIM
1530 aaaaataatcaactggaagacacaggaatagacactataattgg
    KISNWKAPGIDTYG
1485 tataatgggaagagatgtctctatcaagaagataatctctatg
    YYWKMRSSSRERILN
1440 atctttaagatggttaacctccatcaaaactacatctcatagtacac
    IFNEWLNFNQNIPLD
   DOMAIN I
1395 atgtgtaatgtaatggaagacttttaatataatccacagagacac
    MVSGRTLILHKSUDRN
1350 gagttacataattgtctcatatgcagttgcataaatatataatttg
    DVTNYRHISSCTNVIM
   DOMAIN II
1305 aagttttatctctctatatttaaaagaaaaagattcagacagacttt
    KVFSTSILKEKIHRL
1260 aacatgatataatcatctatattaaatcgtgacaccaatttggga
    MNMNESSFKISSNQLG
1215 tgtattgtgcataattcttttagctctaaagaagataatattatgt
    CKLQSLAAKEGIINS
1170 tatattgatgagacatcagaaagaagaaaaatcccaaataattgtg
    YMMKHKQKEEKFYPKYYV
   DOMAIN III
1125 gagttacatttattagcatatagaaaaagcataatcagacagctattttttcacc
    ESYYDIKKAYDTVNH
1080 gaatgttgaattattcataatatttttaatattattagtggtgct
    EWVIESLKYFNYECV
1035 atctatgacatattggatagatgctgtgacagtgaagatatttc
    IIDDIESMMTRWKIF
990 atagctataaatttatataatattcatttatagattatatatatatat
    IGYKFNEYLYGNIKLN
   DOMAIN IV
945 agaggaattttcaagggatatttcttcatcaaatcactctcaatttttttgaagc
    RGILQGDLSNLILLIL
900 attcgatgatgaatatttctcatcaatcagaaagaaattccca
    IQMNVISQIIIEEKFP
855 aaataaccaatcacttttatattgtgatattgagaataagtaca
    KSNHTLYMDLIRM
   DOMAIN V
810 gaaagatagagaaaaagttggaatattccacactgaaattaagaa
    ESRHEMGIINNEIEKE
   DOMAIN VI
Figure: 29. Construction of a continuous ORF in the RT coding region.

A continuous ORF of 464 amino acids is generated in the -3 reading frame after addition of a ‘T’ at 1365 position of the consensus sequence. This region includes the seven conserved RT domains shown in the figure. The inserted nucleotide is coloured red. The seven amino acids deviating from the non-LTR RT consensus sequence are coloured blue. The nucleotide positions are with reference to the consensus sequence given in figure 28.
Figure: 30. Amino acid sequence alignment of EhRLE RT conserved region and Non-LTR RT sequences. The numbers at the end represent the position of the amino acid in the RT ORF. Numbers within the figure represent the number of amino acids present but omitted from the figure. Question marks indicate those instances in which the ends of the ORF have not been determined. An asterix (*) indicates a stop codon at that position. Largely unvaried chemical residues are shown at the top of the alignment. Hydrophobic residue: p, small polar residues: c, charged residues.

For the alignment belong to the following organisms:

to be divergent in EhRLE. The ‘YADD’ box in domain V which is part of the RT active site is mutated to ‘YMDD’ in EhRLE. The coding potential of the RT region was also analysed using a Fourier-based algorithm GeneScan (Tiwari et al., 1997). It can identify a coding region by finding a three-base periodicity observed only in protein coding genes. Since a few mutations do not interfere with the prediction, GeneScan was ideally suited to determine if the RT region is derived from a functional gene or not. A high peak to noise ratio at f=1/3, that is, presence of 3-base correlation suggests that this region is derived from a functional gene (Fig.31). Other peaks are also visible, presumably since the sequence has accumulated mutations, but the peak at f=1/3 is still the highest. In contrast, none of the polyadenylated untranslated RNAs reported from *E. histolytica* (Bhattacharya et al., 2000) show 3-base periodicity by GeneScan analysis.

In order to see the relationship of EhRLE RT with different RT sequences, a phylogenetic tree was drawn with EhRLE RT sequence and RT sequences of several LTR and non-LTR retrotransposons (Fitch and Margolisch, 1967). The data is shown in fig. 32. EhRLE RT groups along with non-LTR RTs and clearly belongs to the non-LTR group of retrotransposons.

**4.2.13 EhRLE of *E. dispar***

In order to find out if EhRLE is unique to *E. histolytica* or it is also present in other species of *Entamoeba*, DNA from the non-pathogenic species *E. dispar* was spotted on a membrane and hybridized at moderate stringency with EhRLE1 probe. There was noticeable signal with *E. dispar*. Other species, such as *Entamoeba moshkovskii* did not give a signal with EhRLE1 (Mittal et al., 1994) suggesting that EhRLE may be present only in *E. histolytica* and *E. dispar* (Fig. 33). In order to compare the sequence of EhRLE from pathogenic *E. histolytica* with it's counterpart in non-pathogenic *E. dispar* (EdRLE), a genomic library of axenic *E. dispar* was searched with EhRLE1 probe (Fig. 34). The library was a kind gift of Dr. Stephan Ortner, Institute for Specific Prophylaxis and Tropical medicine, University of Vienna, Vienna, Austria). Of the positive clones obtained, one was characterized by restriction enzyme digestion and Southern hybridization (Fig. 35). The complete nucleotide sequencing of this clone was done and it was found that the EhRLE-like region is
Figure 31. Fourier analysis of the EhRLE consensus sequence using the genesca software. The genescan pattern of EhRLE consensus sequence shows presence of 3-base correlation suggesting derivation of the region from a functional gene.
23 Populations

Fitch-Margoliash method version 3.573c

\[
\text{Sum of squares} = \sum_{i} \sum_{j} (\text{Obs} - \text{Exp})^2
\]

Negative branch lengths not allowed
remember: this is an unrooted tree!

Sum of squares = 5.39164
Average percent standard deviation = 10.34297
examined 1007 trees

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>And</td>
<td>Length</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.40926</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>0.16697</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.09453</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.11375</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.60631</td>
</tr>
<tr>
<td>3</td>
<td>DONG</td>
<td>0.66462</td>
</tr>
<tr>
<td>3</td>
<td>EhRLE</td>
<td>1.35875</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0.22179</td>
</tr>
<tr>
<td>10</td>
<td>16I</td>
<td>0.91865</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>0.16653</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>0.34106</td>
</tr>
<tr>
<td>14</td>
<td>20Jockey</td>
<td>0.52876</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>0.13271</td>
</tr>
<tr>
<td>13</td>
<td>19G</td>
<td>0.45138</td>
</tr>
<tr>
<td>13</td>
<td>P</td>
<td>0.47956</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>0.06326</td>
</tr>
<tr>
<td>20</td>
<td>T1</td>
<td>1.33365</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>0.12596</td>
</tr>
<tr>
<td>11</td>
<td>Ingi</td>
<td>0.92949</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>0.32223</td>
</tr>
<tr>
<td>5</td>
<td>RIDm</td>
<td>0.95051</td>
</tr>
<tr>
<td>5</td>
<td>RIBm</td>
<td>0.78423</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>0.06789</td>
</tr>
<tr>
<td>18</td>
<td>Txl</td>
<td>1.27495</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>0.03899</td>
</tr>
<tr>
<td>19</td>
<td>Cin4</td>
<td>1.19908</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>0.41538</td>
</tr>
<tr>
<td>7</td>
<td>R2Dm</td>
<td>0.74436</td>
</tr>
<tr>
<td>7</td>
<td>R2Bm</td>
<td>0.73518</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>0.54413</td>
</tr>
<tr>
<td>9</td>
<td>LIHs</td>
<td>0.20006</td>
</tr>
<tr>
<td>9</td>
<td>LIMd</td>
<td>0.19578</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>0.91150</td>
</tr>
<tr>
<td>16</td>
<td>SLACS</td>
<td>0.53468</td>
</tr>
<tr>
<td>16</td>
<td>CRE1</td>
<td>0.76733</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>0.34477</td>
</tr>
<tr>
<td>17</td>
<td>Tyl</td>
<td>1.69196</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0.35082</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>0.23167</td>
</tr>
<tr>
<td>21</td>
<td>COPIA</td>
<td>2.26643</td>
</tr>
<tr>
<td>21</td>
<td>Gypsy</td>
<td>1.00731</td>
</tr>
<tr>
<td>1</td>
<td>Mag</td>
<td>0.68635</td>
</tr>
<tr>
<td>2</td>
<td>DIRS1</td>
<td>0.90681</td>
</tr>
</tbody>
</table>

**Figure 32. Phylogenetic analysis of EhRLE.** The phylogenetic position of EhRLE has been determined. A number of non-LTR and LTR (shown in red) retrotransposon RT sequences have been taken for the analysis. EhRLE groups along with non-LTR retrotransposons. The non-LTR retrotransposons used for the alignment belong to the following organisms: **Dong**-Bombyx mori, **R1Bm**-B. mori, **R1Dm**-Drosophila melanogaster, **LIHs**-human LINE1, **LI Md**-mouse LINE1, **I-D**-melanogaster, **In gi**-T. brucei, **F-D**. melanogaster, **G-D**. melanogaster, **Jockey-D**. melanogaster, **CRE1-C**. fasciculata, **SLACS-T**. brucei, **Tl**-Anopheles gambiae, **Txl-Xaenopus laevis**, **Cin4**-Z. a m (corn). The LTR retrotransposons belong to the following organisms: **DIRS1-D**. discoideum, **Mag-B**. mori, **Copia-D**. melanogaster, **Gypsy-D**. melanogaster, **Tyl-S**. cerevisiae.
Figure: 33. Presence of EhRLE in *E. dispar*. Genomic DNA from *E. histolytica* and *E. dispar* was spotted on gene screen plus membrane. *E. dispar* DNA was spotted in rows 1 and 2 in duplicates ranging from 500 ng to 6.25 ng (columns 1 to 5). *E. histolytica* DNA was spotted in rows 3 and 4 in duplicates ranging from 500 ng to 6.25 pg (columns 1 to 5). EhRLE1 DNA was spotted in rows 5 and 6 in duplicates ranging from 2 ng to 50 pg (columns 1 to 6). Hybridization was done with EhRLE1 insert DNA as probe. *E. dispar* showed the presence of EhRLE but with much lower copy number than *E. histolytica*. 
Figure: 34. Screening of *E. dispar* genomic library with EhRLE1 probe. The genomic library of axenically grown *E. dispar* strain SAW760 was screened with EhRLE1 probe. The primary, secondary and tertiary rounds of screening are shown in panels A, B and C respectively.
Figure: 35. Restriction digestion and Southern hybridization of EdRLE. The Southern hybridization of EdRLE digested with restriction enzymes: Lane 1: PvuII, Lane 2: EcoR1 and NdeI, Lane 3: EcoR1 and XbaI, Lane 4: EcoR1, Lane 5: EcoR1 and Hind III with EhRLE probe at both high stringency (65°C, panel I) and low stringency (50°C, panel II).
Figure: 36. Domain map of EdRLE. The map of EdRLE with respect to EhRLE is shown. The cpn60 and RT regions are indicated in the map. The BLAST results for the different regions are shown in table 6.
present within the clone flanked by sequence with no homology in the databases at the 5' and 3' ends. (Appendix 1). The comparative map of EdRLE with respect to a typical EhRLE is shown in Fig. 36. It is clear that EdRLE has features which are similar to the EhRLEs from *E. histolytica* but it has diverged considerably in terms of nucleotide sequence. The overall sequence identity with EhRLE is about 85%. The RT region also showed comparatively less significant BLASTX scores with known RTs with maximum score = 52 and minimum "p" = 5e-07. The stretch of homologous sequence picked up in the search was much shorter than that picked for EhRLEs. The SOD upstream region is not present in EdRLE. The cpn60 upstream region similar to that of EhRLE is present but with 83.5% homology to the known *E. histolytica* cpn60 gene upstream region. The 5'-end of the cloned fragment has two 42 bp tandem repeats with a single mismatch. The 3' end has no such repeats. BLASTX search results for the RT regions of all EhRLEs and EdRLE are given in Appendix III.
4.3 DISCUSSION

The data presented here clearly indicate that EhRLEs (originally named as HMc) belong to a family of non-LTR retrotransposons which probably have lost their ability to transpose any more. The functions of these elements at present are not known. Similar non-functional retrotransposons are known in almost all eukaryotic organisms (reviewed in Finnegan, 1985, Smit, 1999). In the protozoan parasite *E. histolytica* this is the first report of presence of such an element. Our findings that it is present in *E. dispar* and absent in other species (Mittal et al., 1994) suggest that this must have been acquired after *E. histolytica*/*E. dispar* separated from other *Entamoeba* species.

There are about 140 copies of EhRLE in the genome of *E. histolytica* and these are scattered throughout the genome. The nucleotide sequence identity varied from 96 to 98.1% from the consensus sequence for the different copies. This indicates that these are derived from a single or few copies probably by transposition events. The complete repeating unit has a region that shows high degree of similarity with RT though there was no ORF present. The level of similarity varied among different EhRLE units and probably reflects different rates of accumulation of mutations. It appears from our limited data that the integration site may comprise of 14-mer inverted repeats flanked by a 8-mer direct repeats which have been found at both the ends of one complete repeating unit (EhRLE5) and at one end of another unit (EhRLE4). In general the location of integration is close to genes and there is no specificity involved as different genes have been found associated with different EhRLE units. This is not surprising as genes in *E. histolytica* are packed densely, with the intergenic regions being typically 400 bp to 2.3 kb (Bhattacharya et al., 2000). Even a random transposition event in a genic region will ensure presence of a gene close by.

There are a number of reports that describe retrotransposons where the RT region has lost its ability to encode for a functional enzyme (Fanning et al., 1987, Smit, 1999). Many of these retrotransposons are thought to be immobile at present. It is not yet clear if EhRLE is still capable of transposition. Though none of the EhRLEs analysed by us has an RT region capable of producing a functional enzyme, it is difficult to rule out presence of at least one functional unit among approximately 140
copies that are present in the *E. histolytica* genome. It is also possible that a functional mRNA is generated by post-transcriptional editing, a phenomenon that has been extensively observed with respect to mitochondrial transcripts of many protozoa (Barth., 1999). Though there are a number of nontranslatable polyadenylated RNAs in *E. histolytica* editing has never been observed and the function of these RNA species are not known (Bhattacharya et al., 2000). An attempt was also made to search for probable promoter elements upstream of the putative RT sequences. Other than a modified TATA box no other conserved promoter or enhancer-like element has been observed in these sequences (Fig. 21). Therefore it is rather unlikely that these RT sequences generate functional mRNA. Our attempt to find a functional RT clone from a cDNA library was not successful (data not shown here). However, it is clear that there are no large scale transposition events involving EhRLE taking place in these cells as Southern hybridization patterns of different strains, obtained from different geographical regions, were found to be nearly identical to each other over a period of time (data not shown here).

Reconstruction of DNA transposase sequence has been attempted in the past and active enzyme has been obtained successfully (Ivics et al., 1997). An attempt was made to reconstruct the sequence of the original RT before it started accumulating mutations by comparing RT sequence from different EhRLEs and with a few functional RT sequences. The fact that a large ORF can be obtained by making minimal changes to a consensus sequence, essentially one substitution, and that the predicted amino acid sequence showed higher degree of similarity with functional RTs suggested that these RT-like regions may indeed have originated from functional RTs by accumulation of mutations. In future we do plan to check if this reconstructed protein has any functional activity.

It has been already reported by us that the EhRLE is present in the closely related non-pathogenic species *E. dispar* but absent in other non-pathogenic species of *Entamoeba*, such as *E. invadens* and *E. moshkovskii* (Mittal et al., 1994). The repetitive unit EdRLE from *E. dispar* also has an RT-like region with a maximum of 85% nucleotide sequence identity with one of the EhRLEs. The maximum variation in nucleotide sequence in RT regions of different EhRLEs from the consensus sequence is about 3%. In general average sequence variation among different genes...
between *E. histolytica* and *E. dispar* is around 3% and in intergenic regions average variation is 12% (Willhoeft et al., 1999c). It appears from sequence comparison that the RT region of EdRLE may have been nonfunctional (behaves like an intergenic region) at the time of divergence of *E. histolytica* and *E. dispar*. When RTs from different retrotransposons (LTR and non-LTR) were compared, the *E. histolytica* sequence grouped along with non-LTR retrotransposons suggesting that EhRLE may have been a member of this group. Absence of LTR-like sequences associated with this element indicates further that EhRLE was a non-LTR retrotransposon.

The complete repetitive element also has sequences originally found upstream of other genes, such as SOD gene. The function of this sequence is not known. In full length EhRLE5 and EhRLE4 SOD homology region is found to be close to one end of the repetitive element and upstream of coding regions not part of the element. Since this sequence seems to be present in many locations and associated with genes it is likely to have a regulatory role, such as a general transcription enhancer. Currently there is no evidence in support of this. A stretch of sequence nearly identical to a part of the upstream sequence of the Cpn60 gene is also present in EhRLE. The region of identity is in the 5'-end of the genomic fragment 645 nucleotides upstream of the cpn60 gene (Clark et al., 1995) and is a part of RT coding region of EhRLE. It is likely that Cpn60 gene may be at the boundary of a EhRLE unit and the upstream region of Cpn60 genomic clone may constitute the rest of the repetitive element.

In this report we have described a non-LTR retrotransposon in *E. histolytica* and *E. dispar*. Our data suggest that it is probably immobile right now. The current function of this is not known. Since it is present in large numbers and in all chromosomes these can be used not only as a marker for amebic chromosomes but also in amebic genome mapping studies. An active retrotransposon could be obtained from an inactive one by reconstruction of the RT sequence and by substitution of amino acids (Fig. 29) as reconstruction of a transposase has been successful in fish. Such a reconstructed active retrotransposon can be useful in developing integrative *Entamoeba* vectors for mutagenesis and for transformation. These vectors are needed to carry out functional studies especially in the post-genome era.