RESEARCH ARTICLES

Unexpected Diversity and Differential Success of DNA Transposons in Four Species of *Entamoeba* Protozoans

Ellen J. Pritham,[1] *Cédric Feschotte,*[1] *and Susan R. Wessler*

Department of Plant Biology, The University of Georgia

We report the first comprehensive analysis of transposable element content in the compact genomes (~20 Mb) of four species of *Entamoeba* unicellular protozoans for which draft sequences are now available. *Entamoeba histolytica* and *Entamoeba dispar*, two human parasites, have many retrotransposons, but few DNA transposons. In contrast, the reptile parasite *Entamoeba invadens* and the free-living *Entamoeba moshkovskii* contain few long interspersed elements but harbor diverse and recently amplified populations of DNA transposons. Representatives of three DNA transposase superfamilies (*hobo*/*Activator*/*Tam3*, *Mutator*, and *piggyBac*) were identified for the first time in a protozoan species in addition to a variety of members of a fourth superfamily (Tc1/*mariner*), previously reported only from ciliates and *Trichomonas vaginalis* among protozoans. The diversity of DNA transposons and their differential amplification among closely related species with similar compact genomes are discussed in the context of the biology of *Entamoeba* protozoans.

Introduction

Entamoeba are unicellular eukaryotes that frequently parasitize vertebrate species including human. *Entamoeba histolytica*, the causal factor of amoebic dysentery and liver abscess, leads to the death of up to 100,000 people annually, while *E. dispar* is morphologically identical to *E. histolytica* but does not cause disease (for review, refer Clark (2000)). On an average, the two species display ~95% nucleotide sequence identity in their coding regions and 85% identity in their noncoding intergenic regions (Willhoeft, Buss, and Tannich 2000). Genome sequencing projects are underway for both *E. histolytica* (Loftus et al. 2005) and *E. dispar* as well as for the reptilian parasite *E. invadens* and the free-living *E. moshkovskii*, with the hope that a comparative genomic approach will reveal factors responsible for the pathogenicity of *E. histolytica* and lead to the development of markers for the purpose of species identification. Whole genome sequencing assemblies of the ~20-Mb genomes of *E. dispar* (2- to 2.8-fold coverage), *E. histolytica* (12-fold coverage), and *E. invadens* (2- to 2.8-fold coverage) and 20,000 unassembled shotgun reads for *E. moshkovskii* are publicly available (see *Methods*).

While transposable elements (TEs) make up the largest fraction of the sequenced genomes of multicellular eukaryotes (Kidwell 2002), very little is currently known about the presence and fate of TEs in single-celled eukaryotes. Large-scale sequencing of several protozoan parasites has shown that their genomes are compact (~10–80 Mb; refer Tarleton and Kissinger (2001)) and that the repeat content (0%–14%) is closely correlated with the haploid genome size (Wickstead, Ersfeld, and Gull 2003). In at least one case, the genome appears completely devoid of TEs (e.g., the 22.8-Mb genome of *Plasmodium falciparum*; Gardner et al. (2002)).

TEs can be divided into two major classes according to their mode of transposition. Class 1 elements (or retrotransposons) are mobilized through an RNA intermediate, while Class 2 elements (or DNA transposons) transpose directly via a DNA intermediate. Class 1 elements are traditionally separated in two subclasses, the non–long terminal repeat (LTR) retrotransposons, which include long and short interspersed elements (LINEs and SINEs, respectively) and the LTR retrotransposons. LINEs are extremely widespread and abundant in eukaryotes and have been reported in a number of protozoans, including *E. histolytica* and *E. invadens* (EhLINEs) (Van Dellen et al. 2002). They account for 5% and 0.1% of these genomes, respectively (Van Dellen et al. 2002; Wang et al. 2003). SINEs are noncoding elements (also designated as nonautonomous elements) that replicate by using the machinery encoded by LINEs. SINEs have also been described in *E. histolytica* where they are estimated to make up 1% of the genome (Van Dellen et al. 2002). To date, no LTR retrotransposons have been reported in *Entamoeba* species.

Most DNA transposons are flanked by terminal inverted repeats (TIRs), and in eukaryotes their transposition is characterized by a so-called "cut-and-paste" mechanism whereby the element is excised from one site and reinserted elsewhere in the genome (Craig et al. 2002). Transposition is catalyzed by element-encoded transposase (Tpase) that recognizes and binds at or near the TIRs. DNA transposons are classified into superfamilies based on sequence similarities in the encoded Tpases and other shared structural features, such as the length (and sometimes the sequence) of the target site duplication (TSD) generated upon element insertion and the sequence of the TIRs (Feschotte, Jiang, and Wessler 2002; Robertson 2002). Currently, nine DNA transposon superfamilies are recognized in eukaryotes: CACTA, *hobo*/*Activator*/*Tam3* (*hAT*), *Merlin*/IS*1016*, *Mutator*, *P*-element, PIF/*Harbinger*, *piggyBac*, Tc1/*mariner*, and *Transib* (Feschotte, Jiang, and Wessler 2002; Robertson 2002; Kapitonov and Jurka 2003; Feschotte 2004). Only members of the Tc1/*mariner* superfamily have been previously described in protozoans (Doak et al. 1994; Silva et al. 2005).

To understand more about the distribution and fate of TEs in *Entamoeba*, a computer-based approach was employed to identify all known eukaryotic protein-encoding TEs. Sequences homologous to known Tpase genes were

[1] Present address: Department of Biology, The University of Texas.

Key words: single-celled eukaryote, transposable element, asexual, *Entamoeba*, reverse transcriptase.

E-mail: pritham@uta.edu.

Mol. Biol. Evol. 22(9):1751–1763. 2005
doi:10.1093/molbev/msi169
Advance Access publication May 18, 2005

Table 1
Characteristics of Representative Members of Each *Entamoeba* TE Superfamily

						First TE Hit in BlastP Searches Against GenBank				
Superfamily	Name[a]	Length (bp)	TSD	TIRs (bp)	ORF (aa)[b]	Description[c]	*e* Value	% ID[d]	% Similarity[e]	Length[f]
Retrotransposons										
LINEs (R4-like)	EhLINE1	4,798	Var.	none	1,589	AAA97394 *Ascaris lumbicoides*, R4	8e-35	24	42	726
DNA transposons										
Tc1/*mariner*										
pogo-like	Piglet-Ei1	1,920	TA	23	354	S20478 *Drosophila melanogaster*, *pogo*	9e-24	27	51	278
Fot1-like	*Gemini-Ei1*	2,428	TA	43	617	EAK90938 *Candida albicans, Cirt1*	9e-09	24	41	253
Tc1-like	*Hydargos-Em1*	1,865	TA	23	346	Q16925 *Anopheles albimanus*, *Tucur*	9e-28	29	51	333
Mogwai	*Mogwai-Ei1*	3,528	TA	26	694	P87255 *Aspergillus niger*, *Ant1*	3e-09	29	46	184
Gizmo	*Gizmo-Ei1*	1,691	TA	84	294	O05710 *Anabaena sp.*, IS*An1*	1e-03	22	43	244
Mutator	*MULE-Ei1*	2,882	9-bp	187	456	AAP31248 *Fusarium oxysporum*, *Hop1*	1e-07	21	39	332
hAT	*Chapka-Em1*	2,548	8-bp	21	697	AAD03082 *Bactrocera tryoni*, *Homer*	4e-10	21	39	494
piggyBac	*leapFrog-Ei1*	2,093	TTAA	10	574	AAA87375 *Trichoplusia ni, piggyBac*	1e-34	25	43	549

NOTE.—Var., TSD of variable size.

[a] Name of the representative *Entamoeba* TE, *Eh* stands for *Entamoeba histolytica*, *Ei* for *Entamoeba invadens*, *Em* for *Entamoeba moshkovski*.

[b] Length of the largest protein encoded by *Entamoeba* TE, in amino acids (aa).

[c] GenBank accession number, species, TE name.

[d] Percent identity between *Entamoeba* TE-encoded protein and hit in BlastP alignment.

[e] Percent similarity between *Entamoeba* TE-encoded protein and hit in BlastP alignment.

[f] Length of the alignment produced by BlastP, in amino acids.

identified. The sequences were examined in detail to determine and characterize the TIRs and TSD of at least one copy per superfamily. Using this approach, we identified TEs from four superfamilies: *hAT*, *Mutator*, *piggyBac*, and Tc1/*mariner*. Copy numbers were estimated and compared for each species. These results are discussed in the context of the biology of *Entamoeba*.

Methods

Data Mining

The genome sequencing of four *Entamoeba* genomes has been undertaken in a collaborative effort between the Sanger Institute and The Institute for Genomic Research (TIGR). The sequence reads for each genome are available for Blast search through the Sanger Web site (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/comp_Entamoeba). The contiged genome sequences of *E. histolytica*, *E. dispar*, and *E. invadens* are available through the TIGR Web site (http://www.tigr.org/tdb/e2k1/eha1/). Many *E. histolytica* and *E. invadens* sequence reads have also been deposited in GenBank and can be searched through the National Center for Biotecnology Information (NCBI) Web site (http://www.ncbi.nlm.nih.gov). Queries representing a conserved region of the encoded proteins of all known eukaryotic Class 1 and Class 2 TEs were used in Blast searches (predominantly BlastN and TBlastN) of the *Entamoeba* genome sequences. For LINEs, multiple queries were used that correspond to both the reverse transcriptase (RT) and the endonuclease domains. When possible the initial query corresponded to a PFAM domain found in the conserved domain database at NCBI. To optimize the probability of detection of all protein-encoding TEs within the same species, TBlastN searches were performed iteratively, using first the sequence from a closely related species and then the newly identified sequences against the same genome. The *Entamoeba* sequences that were used as queries are provided as supplementary table 1 (see supplementary Material online). Generally, a hit was considered significant when the *e* value was lower than 10^{-4}. Elements with at least 85% sequence identity were grouped into a family and, when it was possible, assigned to previously established clades based on sequence similarity and phylogenetic analyses of Tpases. (table 1 and figs. 2–4). A canonical TE, which displays all the hallmarks of the superfamily—including a Tpase open reading frame (ORF) of the expected size, TIRs, and TSD, was identified from the contigs of *E. invadens* or assembled from the short sequence reads from *E. moshkovskii*. These canonical TEs have been deposited in Repbase (www.girinst.org/Repbase_Update.html).

Gene Predictions, Phylogenies, and Other Sequence Analysis

Conceptual translations were performed with the Translate program (www.expasy.org/tools/dna.html). The resulting Tpase sequences were aligned with ClustalW (Thompson et al. 1994) using default parameters and edited manually using GeneDoc (K. B. Nicholas, H. B. J. Nicholas, and Deerfield 1997). When necessary, frameshifts were judiciously introduced according to nucleotide alignments of closely related sequences. Phylogenetic trees were generated with MEGA v. 2 (Kumar et al. 2001) using the neighbor-joining method with Poisson correction allowing for multiple substitutions at sites or with PAUP* (Swofford 1999) using the neighbor-joining method with default parameters. In phylogenies with multiple *Entamoeba* Tpase fragments, sequences were included if they necessitated only minor corrections in the ORF and did not have large insertions or deletions.

TE Copy Number Estimates

Copy numbers for each genome were estimated based on the results of Blast searches of the unassembled reads

available at the Sanger Web site. The number of hits could not be directly converted into TE copy number because the Sanger database consists of short unassembled reads (on an average ~650 bp). Copy numbers were estimated based on the effective query length, average length of database reads, and the number of hits and subsequently corrected for the coverage of each genome (see also Zhang and Wessler (2004)). Estimates for the total number of families was based on Blast searches of the contiged genome sequence (TIGR) and therefore represents a minimum number of families.

Results

Class 1 Elements

To identify protein-encoding Class 1 TEs, we utilized both the RT domain (pfam00078) and the integrase (IN) core domain (pfam00665) as initial queries in TBlastN searches. Searches with the RT domain should find virtually all Class 1 protein-encoding TEs, while the presence of the IN domain is characteristic of only LTR retrotransposons. Sequences with significant similarity to the RT domain were identified in all four species of *Entamoeba* but no sequences homologous to the IN domain were identified in any of the *Entamoeba* species examined. These results suggested the absence of complete LTR retrotransposons in these genomes. Consistent with this idea, all *Entamoeba* RT sequences show strongest similarity to those encoded by LINEs of the R4 clade and only weak similarity to those of LTR retrotransposons (data not shown). An endonuclease domain with greatest similarity to those of R4 elements was also detected downstream of most *Entamoeba* RT sequences, a genetic organization typical of the R2 group of LINEs (which includes the R4 clade).

The copy number of LINEs was estimated for each species using both the RT domain and the endonuclease domain (see *Methods*). The number of estimated hits was highest in *E. histolytica* ($n = 293$) and *E. dispar* ($n = 180$), while many fewer hits were obtained in *E. invadens* ($n = 6$) and *E. moshkovskii* ($n = 1$). The number of hits to the R4 endonuclease domain is in agreement with the number of hits to the RT domain, which implies that most *Entamoeba* LINEs contain both RT and endonuclease domains. Together these results suggest that R4-like LINEs are the predominant, if not the only, type of protein-encoding Class 1 element in these species. This is in agreement with previous studies of LINEs in *E. histolytica* and *E. invadens*, which have provided a detailed description of the structure of these elements (Van Dellen et al. 2002; Wang et al. 2003; Mandal et al. 2004). Our results also show that these elements have amplified to different levels in these four species.

Class 2 Elements

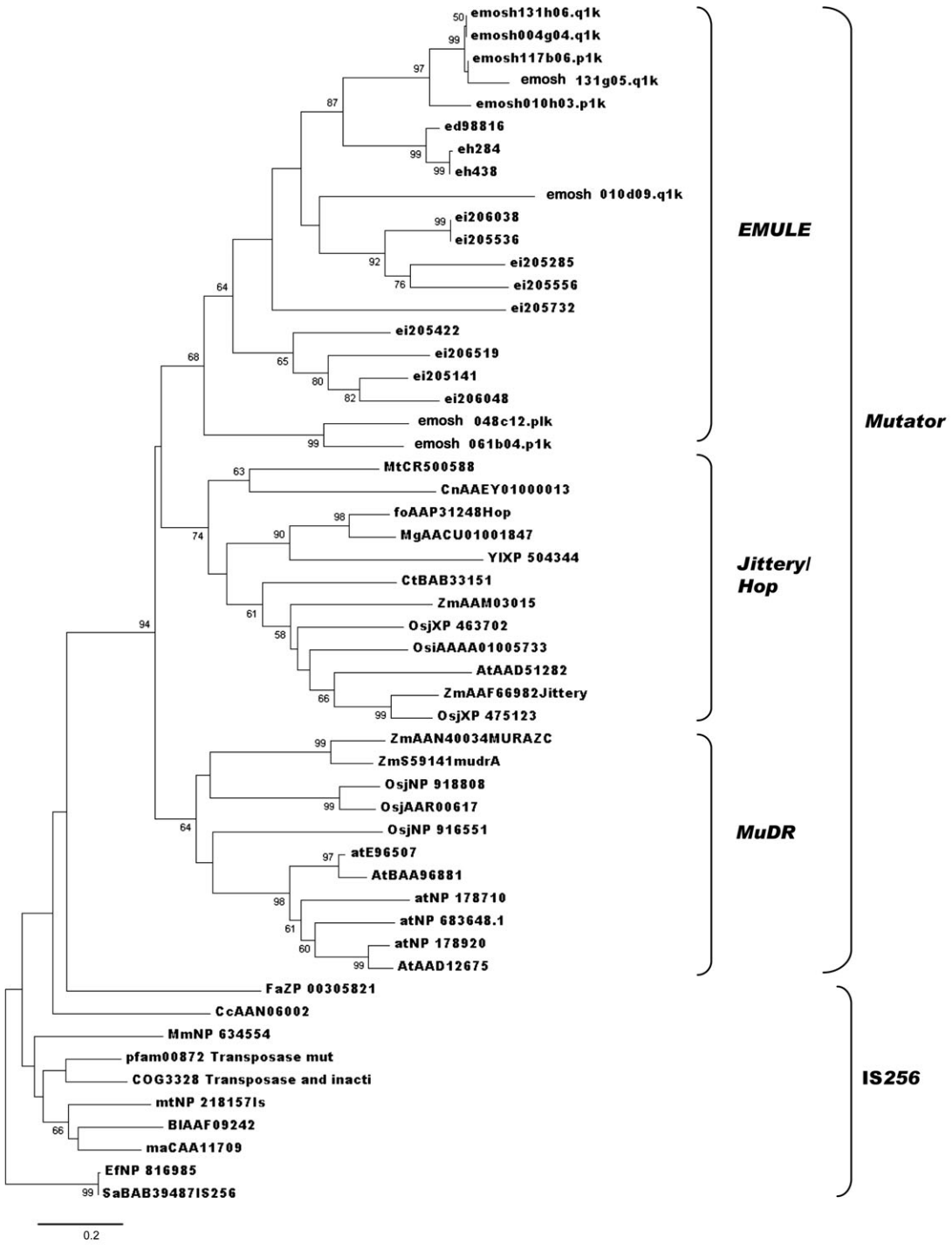
The genomes of the four *Entamoeba* species were searched with TBlastN for sequences homologous to the nine Tpase superfamilies known in eukaryotes. Four superfamilies could be readily detected in at least two of the *Entamoeba* species by using queries representative of each superfamily previously identified in other eukaryotic species (see *Methods*). There were no significant hits (with *e* value $< 10^{-4}$) for the representatives of the remaining five superfamilies. The sections below describe in greater detail the distribution and phylogenetic structure of each superfamily, among the four *Entamoeba* species, as well as the organization and features of at least one canonical member of each superfamily.

Mutator Superfamily

Previously described autonomous *Mutators* range in size from 3 to 5 kb and are characterized by relatively long TIRs (~100–200 bp) and a 9-bp TSD. They encode a putative Tpase of ~700–900 aa that shares weak similarity to those encoded by the IS*256* group of prokaryotic insertion sequences (Eisen, Benito, and Walbot 1994). Three distinct clades of *Mutator* elements were previously recognized in eukaryotes: *Jittery* and *MuDR* in plants and *Hop* in fungi (Lisch et al. 2001; Chalvet et al. 2003; Xu et al. 2004). By using the most conserved region of each clade of *Mutator* Tpase (domain pfam00872) in TBlastN searches, *Mutator*-like Tpase sequences were identified in each of the four *Entamoeba* genomes. The copy numbers vary significantly for each species: 5 in *E. dispar*, 2 in *E. histolytica*, 198 in *E. invadens*, and 322 in *E. moshkovskii*.

Based on Blast scores and sequence comparisons (see below), *Entamoeba Mutator*-like sequences can be divided into two distinct groups. The first group, designated *EMULE*, is represented in each of the four *Entamoeba* species. There are 2 *EMULE* Tpase fragments in *E. dispar*, 5 in *E. histolytica*, 39 in *E. invadens*, and 21 in *E. moshkovskii*. A canonical *EMULE* copy was characterized from the genome of *E. invadens*. This element (*EMULE-Ei1*) is 2,882 bp long and displays structural features typical of plant and fungal *Mutator* elements such as long TIRs (187 bp) and a 9-bp flanking TSD. It contains a single intronless gene that can encode a 456-aa protein that is most similar to the *Hop1* Tpase from *Fusarium oxysporum* (39% similarity, see table 1). A phylogenetic tree was generated from an alignment of 20 *Entamoeba EMULE* putative Tpase domains and 35 *Mutator* Tpase domains from bacteria, archea, fungi, and plants (fig. 1*A*). When the tree is rooted with the bacterial and archeal Tpase fragments, the eukaryotic sequences form a well-supported monophyletic group and can be further divided into three clades, albeit with moderate bootstrap support. *Hop* and *Jittery* form a single clade together apart from the *MuDR* clade, while the *EMULE* sequences form a distinct clade, which itself includes at least five lineages. Two lineages are specific to *E. moshkovskii*, and two lineages are specific to *E. invadens*. The remaining lineage includes sequences from *E. dispar*, *E. histolytica*, and *E. moshkovskii* (fig. 1*A*). The phylogenetic relationships within this lineage are in good agreement with the species phylogenies established from ribosomal sequences (Silberman et al. 1999), suggesting that this lineage of *EMULE* was present in the common ancestor of these species and vertically inherited.

A second, distinct group of *Mutator*-like Tpases is found within the genomes of *E. invadens* and *E. moshkovskii*. The largest of these protein sequences ranges from 350 to 400 aa and displays only weak similarities with the proteins encoded by *EMULE*s and previously described eukaryotic *Mutator* and prokaryotic IS*256* elements. Some of these putative Tpase genes are flanked by long inverted




```
phantom-Ei1: LKMID--ETIQELLNS-DKYYIRDEDNVSLMISREKIDILKNKTIAYLCMDGNFGSCPYD----YEQLYTIHCILTNG--QCF
phantom-Eml: LKRIEVGKRCQEMFEDKEKYFYLNDGENIIFASFEKLKLLQDRSIYYLGMDGTFNSSPKE----FCQLYTIHVILRNG--QCV
MuDR       : IAVMPDSVIEIDVILEDGKYYFSRFFCAFGPCISG----FRDGCRPYLSVDSTALNGRWNG-----HLASATGV--DGHNWMY
IS/256     : LAALFGGT-VGKDTVSRTWRKVKSDWDAWNSRS------LADEPIVRLILDGTVVRVRLDRKATSISLLVVLGVRADG--QKV
pfam00872  : LEELYGGTGVSKSTVSRITKQLDEEVAAYRERP------LEESRYPYLFLDATYVKVREEGRVVSKAVLIALGVTADG--RRE

phantom-Ei1: TIVHSLMVHRREIDYVLLFDINEFFVECCGSFLLNVVVVDFSIKDVSVMVDFETAAINSPSKYECILHGCYF :
phantom-Eml: PLFHVLMKKRTETDYLHIFKVIEDFCLNDLGYSLFN---------KERNVIMDFERAALNAIGCMGCILHGCFF :
MuDR       : PVCFGFFQAETVDNWIWFMKQLKKVVGDMTLLAICSD--------AQKGLVHAVNEVFPYAERREGFRHLMGN :
IS/256     : LLAIKSMGGESAEAWRTVLDDIIKRGLRRPEFLIVDG---------APGLDKAIAVVWDGVPVQRCTVHKHRN :
pfam00872  : ILGIEVGDGESAAFWLTFLDDLKARGLQGVLLVVSDG---------HKGLVAAIRAVFPGASWQRCRVHFLRN :
```

repeats of 180–225 bp, but no TSD was identified. Nonetheless, a central region of ~150 aa can be reasonably well aligned with the most conserved domains of the *Mutator*/IS*256* superfamily of Tpases (fig. 1*B*). Reiterative PSI-Blast searches revealed that these *Entamoeba* proteins also share significant similarity (46%) with Tpase-like sequences described from the insect *Chironomus* and with a variety of animal putative proteins encoded by several uncharacterized mobile elements (unpublished data). It appears that all these putative proteins represent the Tpases of a novel group of DNA transposons distantly related to the *Mutator* superfamily that we called *Phantom*. These elements were not identified in *E. histolytica* or *E. dispar*, yet they represent one of the most abundant group of Tpases found in the genomes of *E. invadens* (~118 copies) and *E. moshkovskii* (~194 copies). Multiple large and nearly identical *Phantom* Tpase fragments were identified in the genome of *E. moshkovskii*, suggesting that this group of elements was recently amplified and may still be active in this species.

Tc1/*mariner* Superfamily

Members of the Tc1/*mariner* superfamily have been identified in a large number of animals, fungi, and plants but only in a few protozoan species. The TBE1 and Tec group of elements were identified in three species of ciliates. These elements are atypical within the superfamily in that they encode multiple putative proteins, one of which is distantly similar to Tc1/*mariner* Tpase. In contrast, a family of elements was recently described in the human parasite *Trichomonas vaginalis* that closely resemble the *mariner*-like elements described in various animals (Silva et al. 2004). Besides the ciliate and *mariner*-like elements, several other large monophyletic subgroups can be distinguished within the Tc1/*mariner* superfamily, notably the Tc1-like, *pogo*-like, and *Fot1*-like groups—none of which have been identified in protists (Shao and Tu 2001; Feschotte and Wessler 2002; Robertson 2002). We searched the four *Entamoeba* genomes with Tpase queries representing each of the known major subgroups of Tc1/*mariner* elements (see *Methods*). A total of 369 and 334 sequences related to Tc1/*mariner* Tpases were found in *E. invadens* and *E. moshkovskii*, respectively, but none could be identified in *E. dispar* and *E. histolytica*. None of these sequences share extensive similarities with the ciliate Tc1/*mariner* Tpases or with any proteins from other protozoans represented in the databases. However, they can be grouped into multiple homogeneous families based on sequence similarities to each other (at least 31 distinct families in *E. invadens*), and most can be readily affiliated with one of the three previously established groups of Tc1/*mariner* elements based on comparisons and phylogenetic analyses of the putative Tpase domains (table 1).

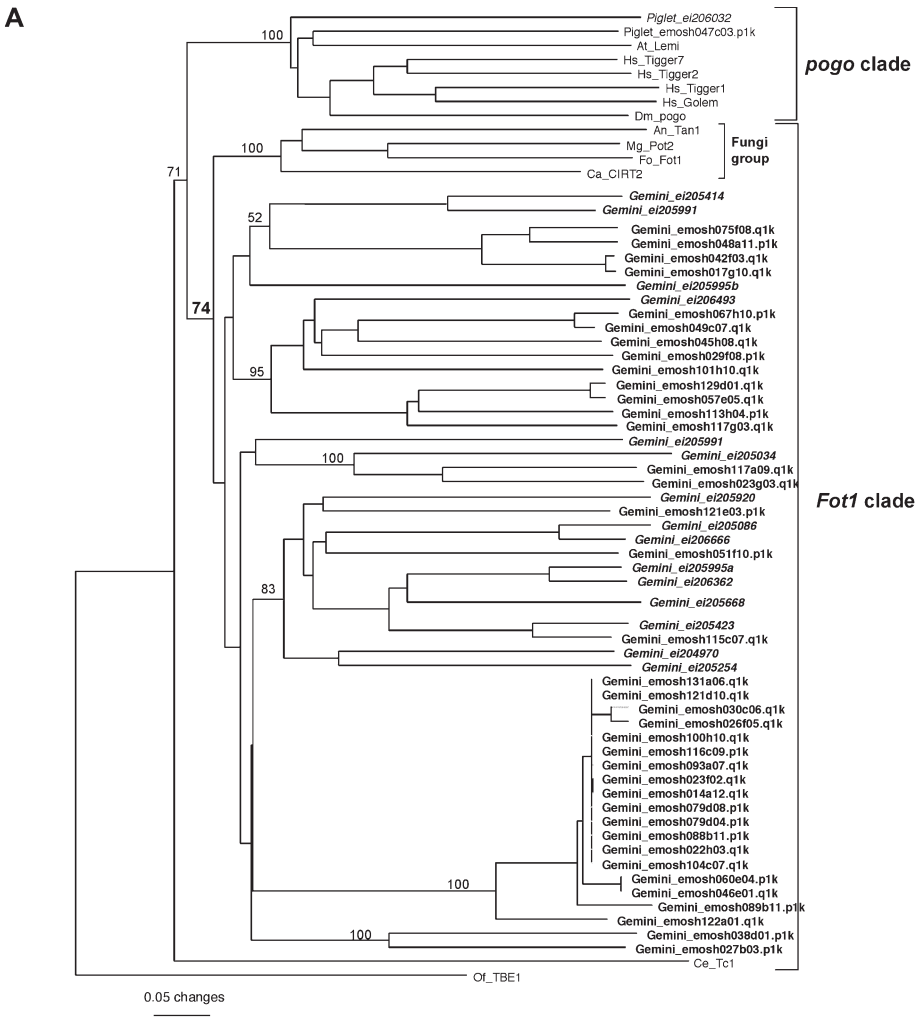
Gemini elements from *E. invadens* and *E. moshkovskii* are most closely related to fungal *Fot1*-like elements (table 1 and fig. 2*A*). *Gemini-Ei1* is an apparently intact and complete member of this group in *E. invadens*. It has 46-bp TIRs that are similar in length and structure to *Fot1* and *Fot1*-like elements (fig. 2*B*) and are flanked by a putative dinucleotide TA TSD. The TIRs of *Gemini* and previously described *Fot1*-like elements have a conserved and peculiar structure consisting of two direct repeats nested within each of the inverted repeats (fig. 2*B*) (Daboussi, Langin, and Brygoo 1992). A single large ORF in *Gemini-Ei1* can potentially encode a 667-aa protein with ~20% identity and ~40% similarity to fungal *Fot1*-like Tpases (table 1). *Gemini* elements have been recently active in *E. moshkovskii* as evidenced by the multiple nearly identical Tpase fragments identified in our search (fig. 2*A*; illustrated by the cluster of fragments with short branches).

Piglet elements from *Entamoeba* belong to the *pogo*-like group of Tc1/*mariner* transposons, which include members in plant, fungi, and animals. *Piglet-Ei1* from *E. invadens* is 1,920 bp long, which fits in the range (2–3 kb) of previously described *pogo*-like elements (Smit and Riggs 1996; Feschotte and Mouchès 2000; Robertson 2002). A neighbor-joining phylogeny also supports the grouping of the *Piglet* Tpase with other *pogo*-like encoded Tpases (fig. 3*A*). *Piglet-Ei1* has 23-bp TIRs with striking similarity to those of other *pogo*-like elements and is flanked by a TA TSD like almost all Tc1/*mariner* superfamily members (fig. 3*B*). *Piglet-Ei1* has coding capacity for a 354-aa protein with 51% similarity to the original *pogo* Tpase from *Drosophila melanogaster* (table 1) (Tudor et al. 1992). *Piglet-Ei1* shares its TIRs with a homogeneous family of short non-coding elements (*mPiglet-Ei1*, ~180 bp) resembling the miniature inverted repeat elements (MITEs) families that are also frequently associated with *pogo*-like transposons in plants and animals (fig. 3*B*) (Feschotte, Zhang, and Wessler 2002).

Hydargos elements are most closely related to the Tc1-like elements previously found in vertebrate and invertebrate species (table 1). *Hydargos-Em1* contains a single ORF encoding a 346-aa protein most similar to *Tucur* from the mosquito *Anopheles albimanus* and *minos* from the medfly *Ceratitis capitata* (table 1 and fig. 4*A*). It is

←

FIG. 1.—Comparative phylogenetic analysis of *Mutator* superfamily proteins and the conserved domain of *phantom*, *MuDR*, and *IS256* Tpases. (*A*) Bootstrapped neighbor-joining tree constructed with MEGA v. 2 from an alignment of a portion of the *Mutator* Tpase protein corresponding to pfam00872 (COG3328) from representative species and rooted with IS*256* proteins from bacteria and archea. Bootstrap values (5,000 replicates) are indicated for the major nodes of the trees. At = *Arabidopsis thaliana*, Bl = *Brevibacterium linens*, Cc = *Clostridium cellulolyticum*, Cn = *Cryptococcus neoformans* var. *neoformans*, Ct = *Carthamus tinctorius*, Ed = *Entamoeba dispar*, Ef = *Enterococcus faecalis V583*, Eh = *Entamoeba histolytica*, Emosh = *Entamoeba moshkovskii*, Ei = *Entamoeba invadens*, Fa = *Ferroplasma acidarmanus*, Fo = *Fusarium oxysporum* f. sp. *melonis*, Ma = *Mycobacterium avium* subsp. *paratuberculosis*, Mg = *Magnaporthe grisea*, Mm = *Methanosarcina mazei Go1*, Mt = *Medicago truncatula*, Osi = *Oryza sativa indica*, Osj = *Oryza sativa japonica*, Sa = *Staphylococcus epidermidis*, Yl = *Yarrowia lipolytica*, Zm = *Zea mays*. *Entamoeba Mutators* (EMULEs), *Hop*/*Jittery*, and *MuDR* form distinct clades in the *Mutator* superfamily phylogeny. (*B*) Clustal alignment of the domain found in the Tpases from the *Mutator*/*IS256* superfamily, corresponding to pfam00872. Ei = *E. invadens* and Em = *E. moshkovskii*, IS/*256* is from *Methylobacterium dichloromethanicum*, and *MuDR* is from *Zea mays*. Residues with related physical or chemical properties were colored in black when they occurred in each sequence and in gray if they occurred in four out of five of the sequences.



B

Tc1/mariner/Fot1			
E. invadens	Gemini1	46	CAGATRSA AGAAWWAKTGGTTGTCAT AGAAAAAGTRGTYGTCAWGA
A. awamori	Tan1	45	ACGTAATCA ACGGTCGRRCGGGCCAC ACGGTCAGGCGGGCMMYCC
F. oxysporum	Fot1	44	AGTCAAGCAC CCATGTAACCACCCCC CCTGGTAACCGACCCMC
M. grisea	Pot2	44	ACGTTGCGTACCCCCTGTTCGGCACCC CCCCTGTTCGGCACCC
B. fuckeliana	Flipper	48	AGTTGAGT ACCCCACTTTCGGACC ACCCCTCTTTTGGACCACCTAAAA
C. albicans	*Cirt1*	37	CCTCACGC CCGGTGGAAATTCTCT CCTGGGAAACTTA

FIG. 2.—Phylogenetic comparison and sequence alignment of *Gemini* and *Fot1* Tpases and TIRs. (*A*) Bootstrapped neighbor-joining phylogenetic tree generated using PAUP* 4.0b8 based on a ClustalX multiple protein alignment with the most conserved domain of the Tc1/*mariner* Tpase (~150 aa) corresponding to the catalytic "DDE" domain. Bootstrap values (1,000 replicates) are indicated for the major nodes of the trees, in particular nodes showing the relationship of *Entamoeba* proteins with the closest known Tpases of other species. Sequences from *Entamoeba moshkovskii* are in bold, and sequences from *Entamoeba invadens* are in bold and italics. Transposon names are followed by "ei" (for *E. invadens*) and "emosh" (for *E. moshkovskii*) and by the contig number. The previously described Tpase sequences are given the corresponding transposon name preceded by the initials of the host species (At: *Arabidopsis thaliana*, Of: *Oxytricha fallax*, An: *Aspergillus nidulans*, Mg: *Magnaporte grisea*, Fo: *Fusarium oxysporum*, Ca: *Candida albicans*, Ce: *Caenorhabditis elegans*, Dm: *Drosophila melanogaster*, Dh: *Drosophila hydei*, Hs: *Homo sapiens*). The tree was rooted with the TBE1 Tpase from the ciliate *O. fallax*. (*B*) A comparison of the consensus TIR sequences from members of the Fot1 and Tc1 clades. The species name is on the left followed by the transposon name, length of TIR, and the consensus TIR sequence. The underlined region represents a direct repeat nested within the inverted repeat.

a 1,861-bp element with 23-bp TIRs flanked by TA TSD (fig. 4*B*). Many *Hydargos* Tpase fragments from the genome of *E. moshkovskii* are nearly identical, suggesting the recent amplification of these lineages in this genome (fig. 4*A*).

Mogwai is a fourth distinct group of Tc1/*mariner* Tpases from *Entamoeba* that share closest similarity to the *Ant1* putative Tpase from the fungus *Aspergillus niger* (table 1). *Ant1* was previously described as a somewhat atypical Tc1/*mariner* element because it is large (~4.8 kb) and appears to have transduced coding and noncoding sequences from the host *amyA* gene (Glayzer et al. 1995). Interestingly, *Mogwai* elements seem to attain a relatively large size (*Mogwai-Ei1* is 3.5 kb) and may also contain transduced host genic sequences in addition or fused to their Tpase coding sequences (data not shown). BlastN searches with the TIRs of *Mogwai-Ei1* revealed several related families of transposons and MITEs in the genome of *E. invadens*.

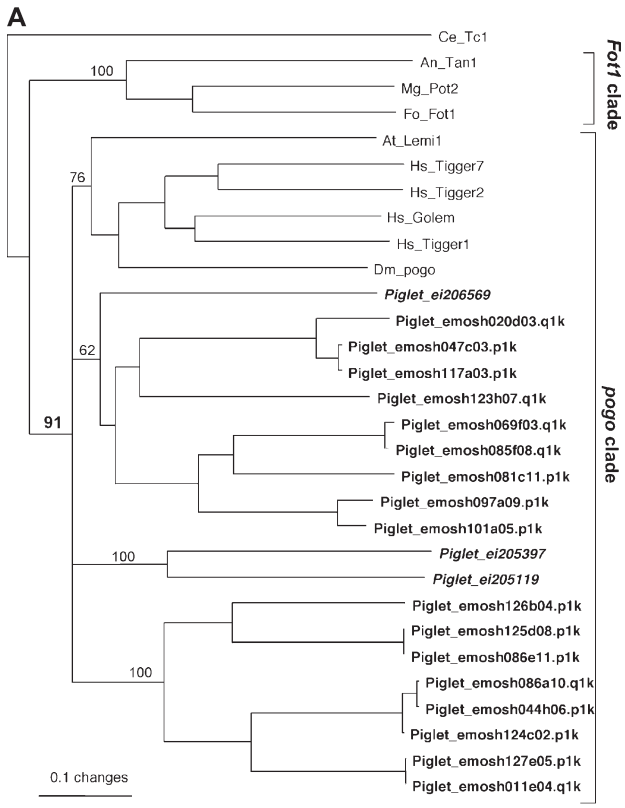


FIG. 3.—Phylogenetic and sequence relationship of *Piglet* with known representative Tpases and TIRs of the Tc1/*mariner* superfamily. (*A*) Trees were generated using PAUP* 4.0b8 with the neighbor-joining method and default parameters based on a ClustalX multiple protein alignment with the most conserved domain of the Tc1/*mariner* Tpase (~150 aa) corresponding to the catalytic "DDE" domain. Bootstrap values (1,000 replicates) are indicated for the major nodes of the trees, in particular nodes showing the relationship of *Entamoeba* proteins with the closest known Tpases of other species. Sequences from *E. moshkovskii* are in bold, and sequences from *E. invadens* are in bold and italics. Transposon names are followed by "ei" (for *E. invadens*) or "emosh" (for *E. moshkovskii*) and by the contig number. Tpase sequences are given the corresponding transposon name preceded by the initials of the host species (At: *Arabidopsis thaliana*, An: *Aspergillus nidulans*, Mg: *Magnaporte grisea*, Fo: *Fusarium oxysporum*, Ce: *C. elegans*, Dm: *Drosophila melanogaster*, Hs: *Homo sapiens*). (*B*) A comparison of the consensus TIR sequences from members of the *pogo* clade. Both the *pogo* and *Piglet* TIRs share sequence similarity, the nucleotides that fit the majority consensus rule are shaded in black.

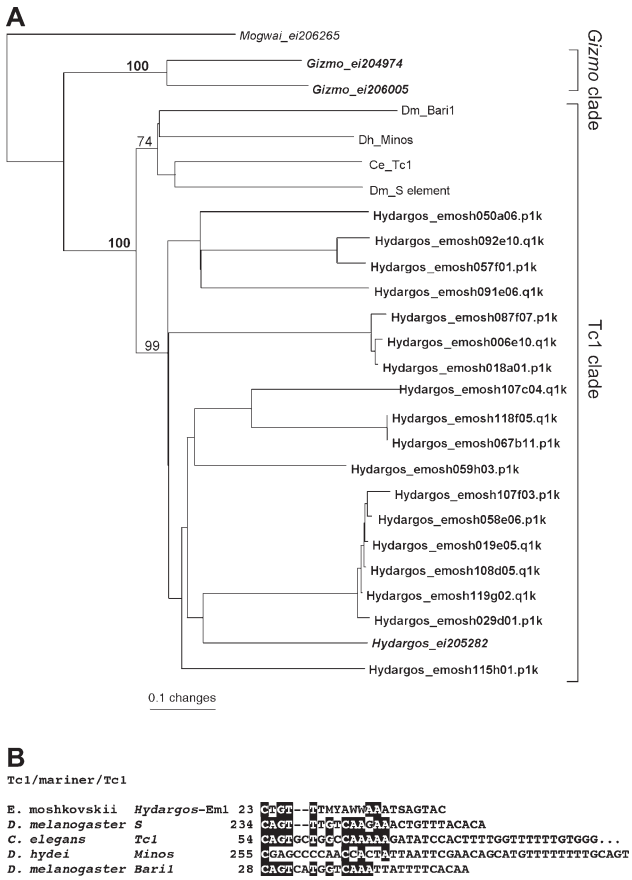


FIG. 4.—Phylogenetic and sequence comparisons of the *Hydargos* Tpases and TIRs from *Entamoeba* group with the members of Tc1 clade of Tc1/*mariner*. (*A*) Trees were generated using PAUP* 4.0b8 with the neighbor-joining method and default parameters based on a ClustalX multiple protein alignment with the most conserved domain of the Tc1/*mariner* Tpase (~150 aa) corresponding to the catalytic "DDE" domain. Bootstrap values (1,000 replicates) are indicated for the major nodes of the trees, in particular nodes showing the relationship of *Entamoeba* proteins with the closest known Tpases of other species. Sequences from *E. moshkovskii* are in bold, and sequences from *E. invadens* are in bold and italics. Transposon names are followed by "ei" (for *E. invadens*) or "emosh" (for *E. moshkovskii*) and by the contig number. Tpase sequences are given the corresponding transposon name preceded by the initials of the host species (Ce: *Caenorhabditis elegans*, Dm: *Drosophila melanogaster*, Dh: *Drosophila hydei*). The tree was rooted with a *Mogwai* Tpase from *E. invadens*. (*B*) A comparison of the consensus TIRs of members of the Tc1 clade. Both Tc1 and the *Hydargos* TIRs share sequence similarity, and the nucleotides that fit the majority consensus rule are shaded in black.

Finally, *E. invadens* and *E. moshkovskii* are hosts to a fifth clade of Tc1/*mariner* elements that encode putative proteins equally distant to other eukaryotic Tc1/*mariner* Tpases. These elements form a monophyletic group that we refer to as *Gizmo*. According to the results of TBlastN searches, *Gizmo* putative Tpases are most similar to those encoded by members of the bacterial IS*630* family, which has been previously affiliated to the eukaryotic Tc1/*mariner* superfamily (43%; see table 1). Alignments of the most conserved region of *Gizmo*- and IS*630*-predicted Tpases show that *Gizmo* proteins possess a conserved D(80)D(32)E motif, which is reminiscent of the catalytic triad found in all Tc1/*mariner* Tpases. *Gizmo-Ei1* is an apparently full-length and intact copy isolated from the genome sequence of *E. invadens*. *Gizmo-Ei1* is a compact transposon (1,691 bp) with relatively long TIRs (84 bp); it is flanked by a TA TSD and has coding capacity for a 294-aa putative Tpase.

hAT Superfamily

Members of the *hAT* superfamily encode a single protein that ranges in size from 500–800 aa, have short TIRs (5-22 bp), and form an 8-bp TSD upon insertion (for recent review, see Kunze and Weil (2002)). Six distinct clades

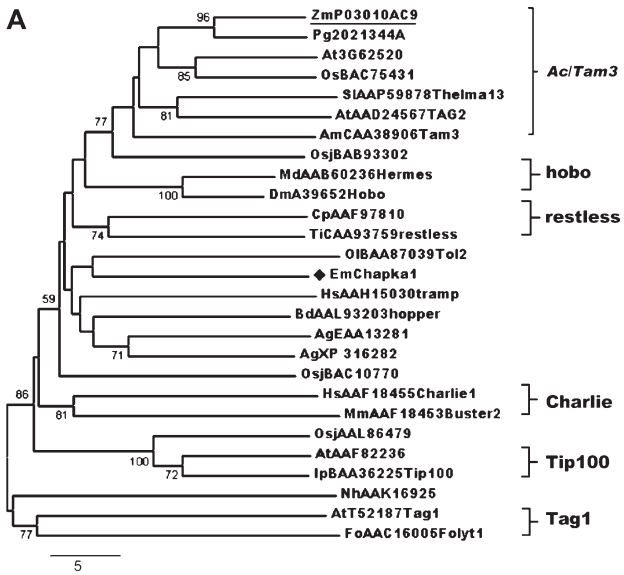


FIG. 5.—Comparative phylogenetic analysis of representative *hAT* superfamily Tpases with the putative *Chapka-Em1* Tpase and comparative sequence analysis of *hAT* and *Chapka-Em1* TIRs. (*A*) Bootstrapped neighbor-joining tree constructed with MEGA v. 2 from an alignment of the complete *hAT* Tpase from representative species and rooted at the midpoint. Zm = *Zea mays*, Pg = *Pennisetum glaucum*, At = *Arabidopsis thaliana*, Os = *Oryza sativa*, Sl = *Silene latifolia*, Am = *Antirrhinum majus*, Md = *Musca domestica*, Dm = *Drosophila melanogastor*, Cp = *Cryphonectria parasitica*, Ti = *Tolypocladium inflatum*, Ol = *Oryzias latipes*, Em = *Entamoeba moshkovskii*, Hs = *Homo sapiens*, Bd = *Bactrocera dorsalis*, Ag = *Anopheles gambiae*, Mm = *Mus musculus*, Ip = *Ipomea purpurea*, Nh = *Nectria haematococca*, Fo = *Fusarium oxysporum*. Only those branches with bootstrap values >50 (averaged from 5,000 bootstrap replicates of the consensus tree) are labeled. Well-supported clades are named for the first described element(s) from the family. *Ac* from *Z. mays* has been underlined. (*B*) Sequence comparison of representative *hAT* TIRs with the TIRs of *Chapka-Em1*. The species name is on the left followed by the name of the transposon, the length of the TIR, and the sequence of the consensus of the TIR. Nucleotides are shaded if they meet the criteria of the majority rule consensus.

have been described (*Ac*/*Tam3*, *hobo*, *restless*, *Charlie*, *Tip100*, and *Tag1*) from plants, animals, and fungi (Robertson 2002) (fig. 5*A*). Representative queries corresponding to the region of the Tpase involved in dimerization (pfam05699) (Essers, Adolphs, and Kunze 2000) were used in the initial searches of *Entamoeba*. *Chapka-Em1*, assembled from short *E. moshkovskii* sequence fragments, is 2,548 bp in length, encodes a single ORF of 697 aa, and harbors perfect 13-bp TIRs that are flanked by an 8-bp TSD. The encoded protein is 39% similar to the Tpase of *homer* from the oriental fruitfly *Bactrocera dorsalis* (table 1). The complete putative Tpase from *Chapka-Em1* was aligned with other representative *hAT* Tpases, and a neighbor-joining phylogeny was constructed (fig. 5*A*). In a midpoint-rooted tree, the putative *Chapka-Em1* Tpase groups with the *hAT*/*Tam3*, *hobo*, *restless*, *Charlie*, and *Tip100* clades, while *Tag1* from *Arabidopsis thaliana* and *Folyt1* from the fungus *F. oxysporum* form a distinct group.

Putative *hAT* Tpase fragments were also isolated from *E. invadens* (43 copies), consisting of at least 11 families, but none of these were identified in either the *E. histolytica* or *E. dispar* genome. Five families were detected in the genome of *E. invadens*, only five were found in the genome of *E. moshkovskii* (~34 copies), and only one (*Chapka*) was found in both species. A sequence alignment was constructed with the TIRs from *Chapka-Em1* and other representative *hAT* TEs (fig. 5*B*). This alignment reveals that the first 8 or 9 bp are well conserved. The *Chapka* TIRs are most similar to TIRs from the *Ac*/*Tam3* clade (*Ac*, *Tag2*, *Tam3*) and to *hobo* from *D. melanogaster*. Thus, the relationship deduced by the comparison of the TIRs is in agreement with the Tpase phylogeny discussed previously.

piggyBac Superfamily

The transposons of this group have a TTAA target site specificity, short TIRs (13–16 bp), and encode a single protein (474–597 aa), which is the Tpase. An active copy has been identified in the lepidopteran, *Trichoplusia ni* (*piggyBac*), and potentially active copies were found in the genome of the dipteran *Anopheles gambiae* (*AgaPB1*) and the crustacean *Daphnia pulcharia* (*Pokey*) (Cary et al. 1989; Penton, Sullender, and Crease 2002; Sarkar et al. 2003). The corresponding Tpases were used as queries in TBlastN searches of the *Entamoeba* genomes and led to the isolation of a 1,958-bp element (*leapFrog-Ei1*) from *E. invadens*. The *leapFrog-Ei1* transposon contains a single ORF of 574 aa that is 43% similar to the original *piggyBac* Tpase and has 10-bp perfect TIRs that are flanked by TTAA (table 1). A portion of the putative Tpases encoded by representative *Entamoeba piggyBac* TEs was aligned with other representative *piggyBac* Tpases, and a neighbor-joining phylogeny was constructed (fig. 6*A*). This phylogeny is rooted with the lepidopterin Tpases *piggyBac* and *yabusame* and illustrates the placement of the *Entamoeba* Tpase fragments within the superfamily. Related Tpase fragments were found in *E. moshkovskii* but were not identified in *E. histolytica* or *E. dispar*.

Blast searches with the TIRs revealed the presence of multiple copies of *leapFrog* in *E. invadens* that are shorter due to deletions within the coding region. A multiple alignment of the *leapFrog-Ei1* TIRs and the TIRs from other representative *piggyBac* TEs revealed strong similarity (fig. 6*B*). The *leapFrog-Ei1* TIRs are most closely related to the *piggyBac* TIRs with 8 out of the 10 bp identical. The identity between the TIRs, the homology with *piggyBac* Tpase and the TTAA target site specificity, supports the grouping of *leapFrog-Ei1* within the *piggyBac* superfamily.

Comparative Analysis of TEs Among *Entamoeba* Species

Copy numbers for each TE family in each *Entamoeba* genome were estimated in order to compare the relative abundance of TEs in each species (fig. 7). LINEs were the predominant component of the genomes of the most closely related species *E. dispar* (~180) and *E. histolytica*

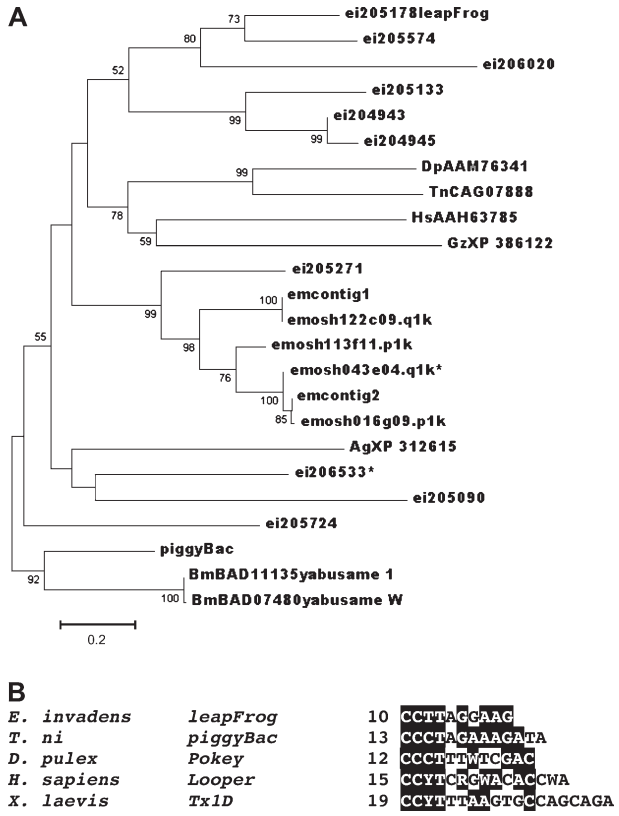


FIG. 6.—Comparative phylogenetic analysis of representative *piggyBac* superfamily Tpases with *Entamoeba* Tpases and sequence comparison of *piggyBac* and *leapFrog-Ei1* TIRs. (*A*) Bootstrapped neighbor-joining tree constructed with MEGA v. 2 from an alignment of a portion of *piggyBac* Tpases from representative species and rooted with the lepidopterin Tpases *piggyBac* and *yabusame*. Ei = *Entamoeba invadens*, Dp = *Daphnia pulicaria*, Tn = *Tetraodon nigroviridis*, Hs = *Homo sapiens*, Gz = *Gibberella zeae*, emcontig = *Entamoeba moshkovskii* assembled from multiple reads, Emosh = *E. moshkovskii* single read, Ag = *Anopheles gambiae*. (*B*) Sequence comparison of representative *piggyBac* TIRs with the TIRs of *leapFrog-Ei1*. The species name is followed by the transposon name, the length of the TIR, and the consensus TIR sequence. Nucleotides are shaded if they meet the criteria of the majority rule consensus.

(293) contributing to ~5% and 7% of the genomes, respectively. The genomes of *E. dispar* and *E. histolytica* are practically devoid of known eukaryotic DNA transposons because only two fragments of *Mutator* Tpase fragments were identified in *E. histolytica* and five fragments were identified in *E. dispar*. The genomes of *E. invadens* and *E. moshkovskii* contained few LINEs, but a rich abundance of DNA transposon families from four superfamilies *hAT*, *Mutator*, Tc1/*mariner*, and *piggyBac* that are estimated to constitute ~7% of the genomes (fig. 7). It is important to note that, as well as reaching high copy numbers, DNA transposons have also greatly diversified in the genomes of *E. invadens* and *E. moshkovskii*. Individual Tpase fragments were assigned to the same family when they shared at least 85% similarity at the amino acid level. Using this criteria, a total of 48 distinct DNA Tpase families were identified in the genome contigs of *E. invadens* (*Mutator* = 8, Tc1/*mariner* = 31, *hAT* = 6, and *piggyBac* = 3). This level of diversification is particularly striking when compared to the 18 families of DNA transposons that are present in the *D. melanogaster* genome (Kaminker et al. 2002).

Discussion

Entamoeba Genomes Harbor Members of Four Eukaryotic DNA Transposon Superfamilies

By exploiting the genomic resources of *E. histolytica*, *E. dispar*, *E. invadens*, and *E. moshkovskii*, we had the unprecedented opportunity to identify and compare the distribution and copy numbers of TE-encoded proteins among four closely related single-celled protozoans. Our analysis led to the identification of members of four Class 2 Tpase superfamiles: *Mutator*, *hAT*, *piggyBac*, and Tc1/*mariner*, none of which were previously known to occur in *Entamoeba*. In addition, this is the first time that members of the *Mutator*, *hAT*, and *piggyBac* superfamilies have been identified in any protozoan species.

To determine if the Tpase fragments that we identified during this study were bonafide TE superfamily members, at least one complete transposon was characterized from each superfamily. These elements not only show significant sequence similarities in their encoded putative proteins with known Tpases of the superfamily but also display striking similarity to other superfamily members in the length and sequence of the TIRs (see table 1 and figs. 2–6). Finally, they are flanked by TSDs whose length and/or sequence are characteristic of the respective superfamilies (TA for Tc1/*mariner*, TTAA for *piggyBac*, 8 bp for *hAT*, and 9 bp for *Mutator*). Therefore, the DNA transposons of *Entamoeba* appear to be typical members of these eukaryotic superfamilies. This is in contrast with the DNA transposons previously described in some other protozoans, which are atypical members of known superfamilies or appear to belong to undescribed groups. These include the TBE1 and Tec elements in ciliates, which are distantly related to Tc1/*mariner* (Doak et al. 1994, 1997), and the yet unclassified Tdd-4 and DDT families in *Dictyostelium discoideum* (Wells 1999; Glockner et al. 2001).

It is also interesting to note that most of the *Entamoeba* DNA transposon families that we analyzed in this study include both Tpase-encoding members and a variety of nonautonomous deletion derivatives. The accumulation and propagation of defective and internally deleted copies is another hallmark of other eukaryotic DNA transposon families (Feschotte, Zhang, and Wessler 2002). In fact, at least two families of *Entamoeba* Tc1/*mariner* elements, *Piglet* and *Mogwai*, are associated with a homogeneous subfamily of short nonautonomous elements that have identical or very similar termini to their Tpase-encoding elements but divergent internal sequences. Although these MITEs are present in relatively moderate copy numbers in *E. invadens*, they have significantly outnumbered their larger presumed partner transposons. This situation is strikingly reminiscent of the amplification of MITEs in many plant and animal genomes, which are also frequently associated with the Tc1/*mariner* superfamily (Feschotte, Zhang, and Wessler 2002). This shows that MITEs can propagate efficiently in a compact protozoan genome.

Finally, we observe that many *Entamoeba* TEs display signs of recent transpositional activity. For example, two copies of *Mutator* from *E. invadens* are 99.9% identical, with only one mismatch over 2,253 bp, but are inserted at different chromosomal positions and are flanked by

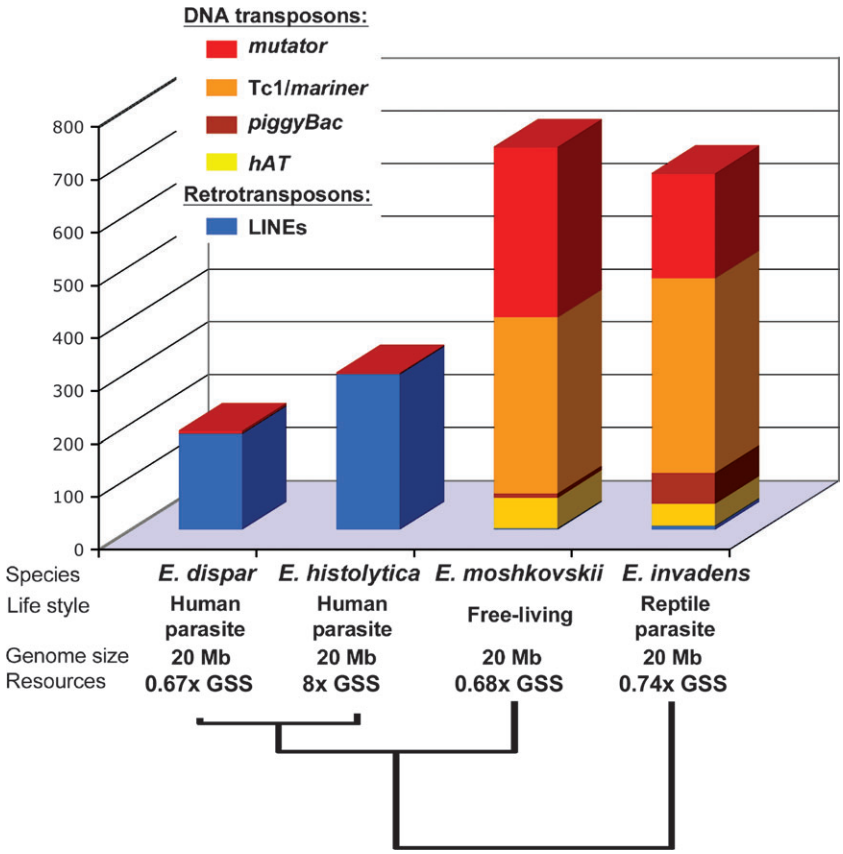


FIG. 7.—Distribution and abundance of DNA transposons and retrotransposons in four species of *Entamoeba* protozoans. TEs were identified and counted as described in the *Methods*. The phylogram below the graph represents the phylogenetic relationship of the four species. The phylogram was constructed with the neighbor-joining method using PAUP* 4.0b8 with default parameters and was based on an alignment of complete 16S ribosomal RNA gene sequences (see also Silberman et al. (1999)) for a more detailed phylogenetic analysis and GenBank accession numbers).

different TSDs, which indicate a very recent transposition event. In addition, several families of the Tc1/*mariner* superfamily include multiple members with very high level of sequence identity in the Tpase (figs. 2–4). Lastly, TE copies harbor long and apparently intact ORFs encoding potentially active enzymes as well as the *cis*-sequences necessary for transposition, such as perfect TIRs. Identifying active endogenous TEs may allow the development of useful molecular tools, such as DNA delivery vectors and mutagenesis systems, for these medically important species and possibly for other eukaryotic pathogens.

Distribution and Diversity of TEs Between *Entamoeba* Species

Both Class 1 and Class 2 elements are detected in all four *Entamoeba* species, but each species has a distinct assortment of elements (fig. 7). LINEs were detected in both *E. dispar* (180 copies) and *E. histolytica* (293 copies), but only few DNA transposons all from the *Mutator* superfamily (five copies and two copies, respectively) were identified in these two species. DNA transposons from four superfamilies were the predominant TEs found in the genomes of *E. invadens* (673 copies) and *E. moshkovskii* (723 copies), while very few LINEs were identified in these species (six copies and one copy, respectively). Previous comparative analyses of closely related species have shown that TE copy numbers can vary by orders of magnitude among species but that overall TE diversity is relatively conserved. For example, the closely related species, *A. thaliana* (125 Mb) and *Brassica oleracea* (~600 Mb), have comparable proportions of Class 1 and Class 2 families, and they share nearly all TE lineages (Zhang and Wessler 2004). However, the copy numbers of TEs from each lineage is almost always greater in *B. oleracea* irrespective of the type of elements, and this increase in copy numbers contributes substantially to the difference in genome size observed between these two plant species. In contrast, while at least one lineage per DNA transposon superfamily in *Entamoeba* is shared between *E. invadens* and *E. moshkovskii*, the majority of the lineages occur in a single *Entamoeba* species. Surprisingly, the differential amplification of retrotransposons and DNA transposons in *Entamoeba* has not led to dramatic changes in genome size (~20 Mb). The uniformity of the genome size among these four *Entamoeba* species coupled with the observed differences in TE diversity and copy numbers suggest that selection is acting to restrict genome size.

It is tempting to speculate that the contrasting assortment and differential success of TEs in *Entamoeba* species reflect differences in their biology and evolutionary histories. *Entamoeba* reproduces using binary fission, and no

sexual stage has been observed in this genus. Theory predicts that DNA transposons will be unable to maintain and propagate in such asexual species (Hickey 1982) because a DNA transposon that is lost by selection, drift, or sequence erosion cannot be reintroduced through sexually mediated genetic exchange. This process could in part explain the dearth of DNA transposons in the genomes of *E. dispar* and *E. histolytica*. However, asexual reproduction cannot be the only factor affecting the DNA transposon diversity in the *Entamoeba* genomes, or we would expect to observe similar patterns of TE diversity in these genomes.

DNA transposons have been observed to move between genomes through nonsexually mediated genetic transfer (i.e., horizontal or lateral transfer) (for example Cary et al. 1989; Daniels et al. 1990; Capy et al. 1998). Horizontal transfer has been hypothesized previously to explain the identification of Tc1/*mariner* Tpase fragments in the genome of the asexual bdelloid rotifers (Arkhipova and Meselson 2000). Although horizontal introduction of *Entamoeba* TEs cannot be ruled out at this point, several lines of evidence support their vertical transmission through the evolution of the four species studied. First, LINE-like retrotransposons and *Mutator* DNA transposons are found in each *Entamoeba* species, a pattern that suggests that related TEs were present in their last common ancestor. Second, lineages of *hAT* (*Chapka*), *Mutator* (*EMULE*), and Tc1/*mariner* (*Piglet*, *Gemini*, and *Hydargos*) contain Tpase fragments from two or more *Entamoeba* species, and their phylogenetic relationship is consistent with the species phylogeny (figs. 1, 2, 3, 4 and data not shown). This is a strong indication that these lineages were present in the common ancestor and were vertically inherited. Third, *E. invadens* and *E. moshkovskii* have a similar distribution of TE families, despite the fact that *E. moshkovskii* is most closely related to *E. dispar* and *E. histolytica*. (see phylogram on fig. 7; Silberman et al. (1999)). These data and the presence of shared TE lineages between *E. invadens* and *E. moshkovskii* therefore suggest a scenario where all types of TEs (both Class 1 and the various Class 2 superfamilies) were present in the common ancestor of these four species but that TE diversity was later dramatically decreased in *E. dispar* and *E. histolytica* or their common ancestor.

How can the decreased diversity of TEs in *E. dispar* and *E. histolytica* be explained? It has been suggested that the maintenance of TEs is dependent on the effective population size of the host species and differs for Class 1 and Class 2 elements due to the differences in their transposition mechanism (see Lynch and Conery (2003)). Because *E. dispar* and *E. histolytica* are human parasites, like *P. falciparum*, the effective population size is influenced by that of their human host. The near absence of DNA transposons might be explained if the effective population size of *E. dispar* and *E. histolytica* is below that sufficient for the proliferation of DNA transposons. By the same token, because the diversity of DNA transposons has been maintained in the genomes of *E. invadens* and *E. moshkovskii*, their effective population size might therefore be larger. Previous studies have shown that *E. dispar* and *E. histolytica* have gone through recent population bottlenecks (Ghosh et al. 2000). In *P. falciparum*, it has been speculated that severe bottlenecks have greatly diminished its genetic diversity (Rich et al. 1998) and this species is apparently devoid of all Class 1 and Class 2 TEs (Gardner et al. 2002). Analyzing the TE content of other *Entamoeba* or related species may reveal whether these patterns are indicative of particular biological and epidemiological features and how TEs may have contributed to the structure and dynamics of these compact eukaryotic genomes.

Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank TIGR and The Wellcome Trust Sanger Institute for making the *Entamoeba* genomic sequence available to the public. The Biotechnology and Biological Sciences Research Council has funded the Sanger Institute Pathogen Sequencing Unit, in collaboration with Graham Clark at the London School of Hygiene and Tropical Medicine, to undertake whole genome shotgun sequencing of *E. invadens*, *E. dispar*, *E.terrapinae*, and *E. moshkovskii*. We would like to thank Dr. Clark for permission to use the Sanger-generated *Entamoeba* sequence for the purposes of this study. The TIGR sequencing effort is supported by an award from the National Institute of Allergy and Infectious Diseases. This work was supported by a postdoctoral research fellowship in Biological Informatics from the National Science Foundation 0107590 to E.J.P. and a grant from the National Science Foundation Plant Genome program to S.R.W.

Literature Cited

Arkhipova, I., and M. Meselson. 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. USA **97**:14473–14477.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1998. Dynamics and evolution of transposable elements. Springer-Verlag, Austin, Tex.

Cary, L. C., M. Goebel, B. G. Corsaro, H. G. Wang, E. Rosen, and M. J. Fraser. 1989. Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology **172**:156–169.

Chalvet, F., C. Grimaldi, F. Kaper, T. Langin, and M. J. Daboussi. 2003. Hop, an active Mutator-like element in the genome of the fungus Fusarium oxysporum. Mol. Biol. Evol. **20**:1362–1375.

Clark, C. G. 2000. The evolution of *Entamoeba*, a cautionary tale. Res. Microbiol. **151**:599–603.

Craig, N. L., R. Craigie, M. Gellert, and A. M. Lambowitz. 2002. Mobile DNA II. American Society for Microbiology Press, Washington, D. C.

Daboussi, M.-J., T. Langin, and Y. Brygoo. 1992. Fot1, a new family of fungal transposable elements. Mol. Gen. Genet. **232**:12–16.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. 1990. Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics **124**:339–355.

Doak, T. G., F. P. Doerder, C. L. Jahn, and G. Herrick. 1994. A proposed superfamily of transposase genes: transposon-like

elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA **91**:942–946.

Doak, T. G., D. J. Witherspoon, F. P. Doerder, K. Williams, and G. Herrick. 1997. Conserved features of TBE1 transposons in ciliated protozoa. Genetica **101**:75–86.

Eisen, J. A., M. I. Benito, and V. Walbot. 1994. Sequence similarity of putative transposases links the maize Mutator autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res. **22**:2634–2636.

Essers, L., R. H. Adolphs, and R. Kunze. 2000. A highly conserved domain of the maize activator transposase is involved in dimerization. Plant Cell **12**:211–224.

Feschotte, C. 2004. Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Mol. Biol. Evol. **21**:1769–1780.

Feschotte, C., N. Jiang, and S. R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3**:329–341.

Feschotte, C., and C. Mouchès. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol. Biol. Evol. **17**:730–737.

Feschotte, C., and S. R. Wessler. 2002. *Mariner*-like transposases are widespread and diverse in flowering plants. Proc. Natl. Acad. Sci. USA **99**:280–285.

Feschotte, C., X. Zhang, and S. Wessler. 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. Pp. 1147–1158 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, D. C.

Gardner, M. J., N. Hall, E. Fung et al. (42 co-authors). 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature **419**:498–511.

Ghosh, S., M. Frisardi, L. Ramirez-Avila et al. (8 co-authors). 2000. Molecular epidemiology of *Entamoeba* spp.: evidence of a bottleneck (Demographic sweep) and transcontinental spread of diploid parasites. J. Clin. Microbiol. **38**:3815–3821.

Glayzer, D. C., I. N. Roberts, D. B. Archer, and R. P. Oliver. 1995. The isolation of Ant1, a transposable element from Aspergillus niger. Mol. Gen. Genet. **249**:432–438.

Glockner, G., K. Szafranski, T. Winckler, T. Dingermann, M. A. Quail, E. Cox, L. Eichinger, A. A. Noegel, and A. Rosenthal. 2001. The complex repeats of Dictyostelium discoideum. Genome Res. **11**:585–594.

Hickey, D. A. 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics **101**:519–531.

Kaminker, J. S., C. M. Bergman, B. Kronmiller et al. (9 co-authors). 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. **3**:RESEARCH0084.

Kapitonov, V. V., and J. Jurka. 2003. Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc. Natl. Acad. Sci. USA **100**:6569–6574.

Kidwell, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica **115**:49–63.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Arizona State University, Tempe, Ariz.

Kunze, R., and C. F. Weil. 2002. The hAT and CACTA superfamilies of plant transposons. Pp. 565–610 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, D. C.

Lisch, D. R., M. Freeling, R. J. Langham, and M. Y. Choy. 2001. Mutator transposase is widespread in the grasses. Plant Physiol. **125**:1293–1303.

Loftus, B., I. Anderson, R. Davies et al. (51 co-authors). 2005. The genome of the protist parasite *Entamoeba histolytica*. Nature **433**:865–868.

Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science **302**:1401–1404.

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

Nicholas, K. B., H. B. J. Nicholas, and D. W. I. Deerfield. 1997. GeneDoc: analysis and visualization of genetic variation. EMBNEW **4**:14.

Penton, E. H., B. W. Sullender, and T. J. Crease. 2002. Pokey, a new DNA transposon in Daphnia (Cladocera: Crustacea). J. Mol. Evol. **55**:664–673.

Rich, S. M., M. C. Licht, R. R. Hudson, and F. J. Ayala. 1998. Malaria's eve: evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum. Proc. Natl. Acad. Sci. USA **95**:4425–4430.

Robertson, H. M. 2002. Evolution of DNA transposons. Pp. 1093–1110 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, D.C.

Sarkar, A., C. Sim, Y. S. Hong, J. R. Hogan, M. J. Fraser, H. M. Robertson, and F. H. Collins. 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Mol. Genet. Genomics **270**:173–180.

Shao, H., and Z. Tu. 2001. Expanding the diversity of the IS630-Tc1-*mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics **159**:1103–1115.

Silberman, J. D., C. G. Clark, L. S. Diamond, and M. L. Sogin. 1999. Phylogeny of the genera *Entamoeba* and *Endolimax* as deduced from small-subunit ribosomal RNA sequences. Mol. Biol. Evol. **16**:1740–1751.

Silva, J. C., F. Bastida, S. L. Bidwell, P. J. Johnson, and J. M. Carlton. 2005. A potentially functional mariner transposable element in the protist Trichomonas vaginalis. Mol. Biol. Evol. **22**:126–134.

Smit, A. F. A., and A. D. Riggs. 1996. *Tiggers* and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93**:1443–1448.

Swofford, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, Mass.

Tarleton, R. L., and J. Kissinger. 2001. Parasite genomics: current status and future prospects. Curr. Opin. Immunol. **13**:395–402.

Thompson, J. D., D. Desmond, D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tudor, M., M. Lobocka, M. Goodwell, J. Pettitt, and K. O'Hare. 1992. The *pogo* transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet. **232**:126–134.

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

Wang, Z., J. Samuelson, C. G. Clark, D. Eichinger, J. Paul, K. Van Dellen, N. Hall, I. Anderson, and B. Loftus. 2003. Gene discovery in the *Entamoeba invadens* genome. Mol. Biochem. Parasitol. **129**:23–31.

Wells, D. J. 1999. Tdd-4, a DNA transposon of Dictyostelium that encodes proteins similar to LTR retroelement integrases. Nucleic Acids Res. **27**:2408–2415.

Wickstead, B., K. Ersfeld, and K. Gull. 2003. Repetitive elements in genomes of parasitic protozoa. Microbiol. Mol. Biol. Rev. **67**:360–375.

Willhoeft, U., H. Buss, and E. Tannich. 2000. Genetic differences between *Entamoeba histolytica* and *Entamoeba dispar*. Arch. Med. Res. **31**:S254.

Xu, Z., X. Yan, S. Maurais, H. Fu, D. G. O'Brien, J. Mottinger, and H. K. Dooner. 2004. Jittery, a Mutator distant relative with a paradoxical mobile behavior: excision without reinsertion. Plant Cell **16**:1105–1114.

Zhang, X., and S. R. Wessler. 2004. Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. Proc. Natl. Acad. Sci. USA **101**:5589–5594.

Laura Katz, Associate Editor

Accepted May 12, 2005