

ARTICLE

Received 24 Jan 2012 | Accepted 19 Apr 2012 | Published 22 May 2012

DOI:10.1038/ncomms1855

Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*

Vijay Pal Yadav¹, Prabhat Kumar Mandal^{2,†}, Alok Bhattacharya² & Sudha Bhattacharya¹

Non-long terminal repeat Retrotransposons are referred to as long interspersed nuclear elements (LINEs) and their non-autonomous partners are short interspersed nuclear elements (SINEs). It is believed that an active SINE copy, upon retrotransposition, generates near identical copies of itself, which subsequently accumulate mutations resulting in sequence polymorphism. Here we show that when a retrotransposition-competent cell line of the parasitic protist *Entamoeba histolytica*, transfected with a marked SINE copy, is induced to retrotranspose, >20% of the newly retrotransposed copies are neither identical to the marked SINE nor to the mobilized resident SINEs. Rather they are recombinants of resident SINEs and the marked SINE. They are a consequence of retrotransposition and not DNA recombination, as they are absent in cells not expressing the retrotransposition functions. This high-frequency recombination provides a new explanation for the existence of mosaic SINEs, which may impact on genetic analysis of SINE lineages, and measurement of phylogenetic distances.

¹ School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India. ² School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India. [†]Present address: McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA. Correspondence and requests for materials should be addressed to S.B. (email: sbjnu10@gmail.com or sb@mail.jnu.ac.in).

Retrotransposons of the non-long terminal repeat (non-LTR) category are generally referred to as long interspersed nuclear elements (LINEs) and their non-autonomous partners are called short interspersed nuclear elements (SINEs)¹. SINEs are extremely efficient genome invaders. The most well-known SINE, the human Alu element is found in 1.1 million copies in the human genome². Individual SINE copies typically show extensive sequence polymorphism^{3–5}. How this polymorphism is generated is not clear, and direct sequence comparison of newly retrotransposed SINE copies with their 'source' copy has not been reported. It should be possible to address this issue using retrotransposition-competent cell lines. *In vivo* retrotransposition of the human LINE, L1⁶ in cultured cells, and SINEs, Alu⁷ and SVA^{8,9}; and a LINE/SINE pair from eel¹⁰, have been well documented. However, analysis of sequence changes in retrotransposed copies compared with the marked SINE copy after a round of retrotransposition has been limited.

We have been studying the LINE and SINE elements in the early branching parasitic protist *Entamoeba histolytica*. These retrotransposons comprise ~11% of the 23-Mb genome^{11–13}. EhLINEs belong to the R2 group of non-LTR retrotransposons^{14–16}. EhLINEs/SINEs are generally located in intergenic regions^{11,13}, but not within genes. The 4.8-kb EhLINE1 contains two open reading frames (ORFs) (Fig. 1a). ORF2 encodes the reverse transcriptase (RT) and endonuclease (EN) activities typically required for non-LTR element retrotransposition^{6,17–19}. EhSINE1 (550 bp) is the likely non-autonomous partner of EhLINE1. The two share a 78-bp stretch of sequence homology at their 3' ends (Fig. 1a)^{10,11,20}. Of the 742 EhLINE1 copies in the *E. histolytica* genome, 88 are full-length, but they lack complete ORFs¹³. Therefore to study retrotransposition in these cells it is necessary to generate a cell line that expresses both ORFs of EhLINE1. Here we report the construction of such a cell line, which expresses ORF2 in a tetracycline (tet)-inducible manner, and can retrotranspose a SINE copy in the presence of tet. This is the first report of active retrotransposition in a parasitic protist cell line. Using this system, we demonstrate a novel feature of high-frequency recombination between SINE copies during retrotransposition.

Results

A retrotransposition-competent *E. histolytica* cell-line. Although EhSINE1 transcripts are abundantly present in *E. histolytica* cells^{21,22}, full-length transcripts of EhLINE1 are not detected (Fig. 1b). We show that *E. histolytica* cells maintained in the lab express ORF1p (Fig. 1c), but fail to express detectable levels of ORF2p; hence these cells are not expected to be retrotransposition competent.

The study of active retrotransposition requires the construction of a cell line expressing the LINE-encoded ORFs, as achieved with human L1⁶, which could also retrotranspose SINEs, such as Alu⁷ and SVA^{8,9}. To express the functions required for retrotransposition, we reconstructed the complete ORF2 (lacking any stop codons) by overlapping PCR (Fig. 2a), cloned it in a tet-inducible expression vector (Fig. 2b) and introduced it into *E. histolytica* cells to obtain the cell line Eh-ORF2. Our strategy to measure retrotransposition was to introduce in this cell line a plasmid containing a marked EhSINE1 copy (with a 25-bp GC-rich tag), and a known target site of EhSINE1 insertion¹⁵ (Fig. 2c). The sequence used as the target site for insertion was identified in a previous study in which we had looked for 'empty' sites in the genome, where a SINE element was missing in one chromosomal copy, but had inserted in another copy of the *E. histolytica* polyploid genome. The empty versus occupied site was differentiated by PCR amplification with flanking primers. The target site used in this assay was one such sequence with an empty site¹⁵. Hence, it is a sequence used by the *E. histolytica* retrotransposition machinery for SINE insertion *in vivo*.

Retrotransposition events occurring at this target site were scored without using selection pressure. Being PCR-based, we

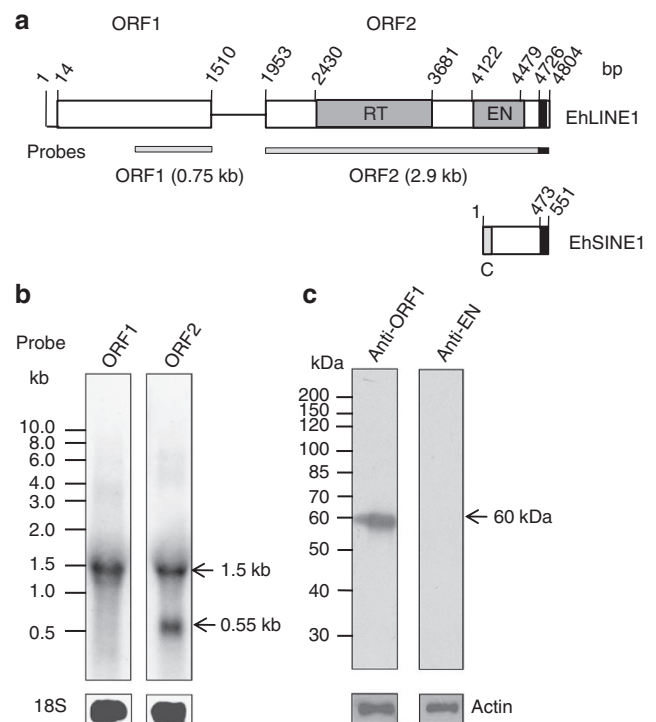


Figure 1 | Expression of EhLINE1 and EhSINE1 in normally proliferating *E. histolytica*. (a) Organization of EhLINE1 and EhSINE1¹¹. The ORF2 contains RT and EN domains¹⁵. The sequence similarity of 78 bp between EhLINE1 and EhSINE1 at their 3' ends is shown by a solid black bar. The region marked 'C' (100 bp) in EhSINE1 is conserved in all EhSINEs and may contain the SINE promoter¹². (b) Northern hybridization was performed with total cellular RNA isolated from normally proliferating trophozoites with the hybridization probes as indicated in (a). The 0.55-kb band corresponds to EhSINE1 and is seen with ORF2 probe containing the 78-bp common 3'-end. No full-length EhLINE1 transcripts of 4.8 kb are seen. Instead the ORF1 and 2 probes hybridize with a 1.5-kb band, which probably corresponds to transcripts from the abundant truncated EhLINE copies¹¹. (c) Western blotting followed by immunodetection with affinity-purified anti-ORF1 and anti-EN antibodies was performed using total cellular lysate. Band of 60 kDa (ORF1p) is indicated. Anti-actin antibody was used for loading control.

expect the scoring to be very sensitive. The doubly transfected cell line (Eh-ORF2-SN) expressed 2.9 kb ORF2 transcript (Fig. 3a) and ORF2p (111 kDa) (Fig. 3b) in a tet-inducible manner. The constitutive expression of ORF1p (60 kDa) was unaltered on tet induction (Fig. 3b). This cell line also expressed the transcript corresponding to the marked SINE1 (Fig. 3c). In the presence of tet, this cell line is expected to be retrotransposition competent. We added tet to cultures in early log phase and harvested the cells after 48 h (late log). We scored retrotransposition of the marked SINE copy to the insertion target site by PCR amplification of total genomic DNA using two sets of primer pairs to discount the possibility of PCR artefacts (Fig. 4). The identity of the amplicons was further confirmed by Southern hybridization with the marked SINE probe. We did not obtain amplicons in the absence of tet, whereas in the presence of tet, specific amplicons expected from the mobilization of the marked SINE to the insertion site were obtained with both primer pairs (Fig. 4a,b). We also did not obtain amplicons when tet was added to a cell line containing the marked SINE and insertion hotspot but lacking ORF2 (Eh-SN). A hallmark of retrotransposition is the generation of target site duplications (TSD). We induced retrotransposition by tet addition in three independently grown cultures and

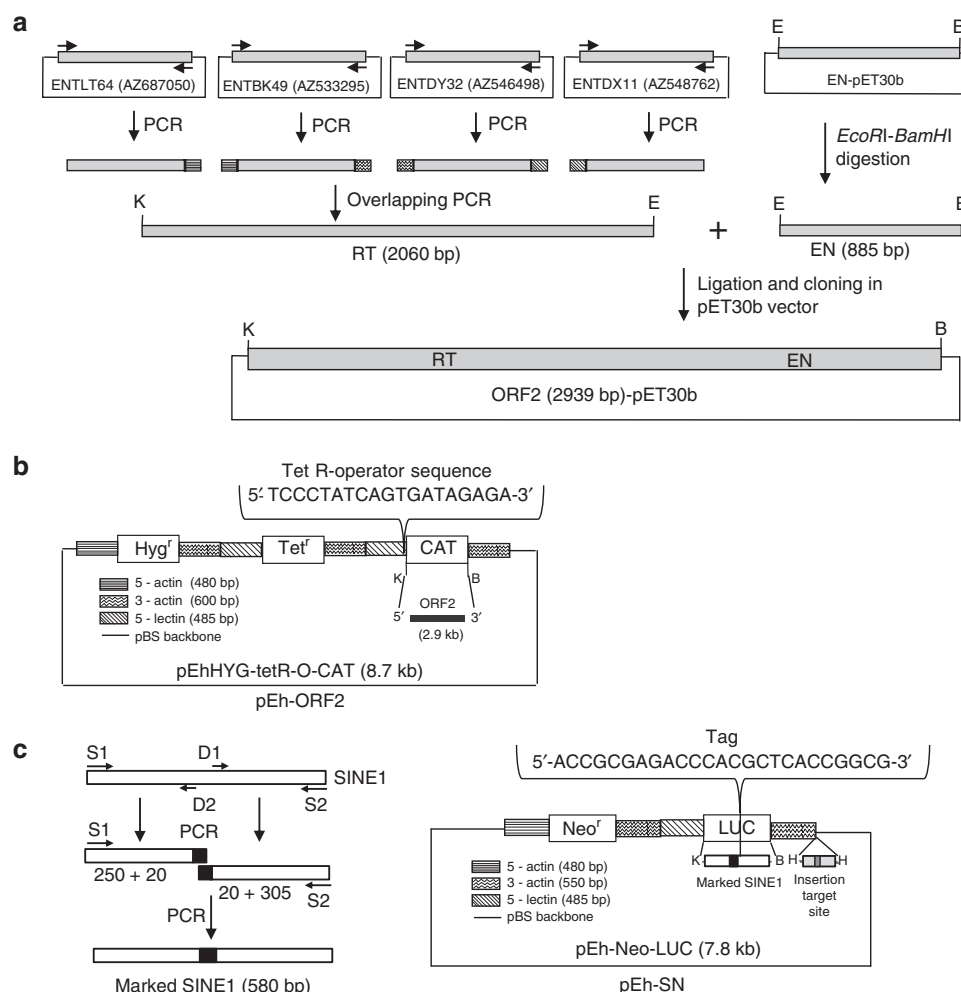


Figure 2 | Construction of complete ORF2 and pEh-ORF2 and pEh-SN plasmids. (a) Overlapping GSS clones having maximum similarity with consensus sequence of ORF2 of EhLINE1 and lacking any stop codons were selected for reconstruction of ORF2. PCR-amplified fragments of the indicated GSS clones were used in an overlapping PCR to obtain 2,060 bp fragment containing the RT domain, which was later stitched with previously cloned fragment containing the EN domain¹⁵ to obtain DNA sequence of the full-length ORF2. It was cloned in pET30b vector at *KpnI*-*Bam*HI sites. Accession numbers of GSS clones are given in brackets K-*KpnI*, B-*Bam*HI, E-*Eco*RI. **(b)** Cloning of ORF2 in place of CAT in the tet-inducible vector pEhHYG-tetR-O-CAT at *KpnI* and *Bam*HI sites to get pEh-ORF2. The 5'- and 3'-actin and 5'-lectin sequences contain regulatory sequences from the *E. histolytica* genes to drive transcription. A sequence of 19 bp inserted between TATA box and ATTCA initiator element in lectin promoter, which acts as a TetR-operator is shown³⁰. **(c)** Cloning of marked EhSINE1 and insertion target site in *E. histolytica* vector pEh-Neo-LUC. A 25-bp GC-rich DNA tag (with no match with *E. histolytica* genome) was inserted at position 250 from 5'-end in EhSINE1 (555 bp). To insert this tag, the EhSINE1 was first PCR amplified in two fragments, that is, 270 and 325 bp with primer pairs S1-D2 and D1-S2, respectively. The primers D2 and D1 contained an overhang of 20 nucleotides each at their 5'-end corresponding to the tag sequence and had 15 nucleotides complementary to each other. These two fragments were used as templates in a second PCR with primers S1-S2 to yield the 580-bp marked EhSINE1 with 25 bp GC-rich tag. This was cloned in *KpnI* and *Bam*HI restriction sites by replacing the 1.65-kb LUC. A genomic target site for EhSINE insertion was also provided by inserting the 176-bp DNA fragment containing this site, downstream to 3'-actin sequence at the *Hind*III (H) site. This site corresponds to the endonuclease nicking hot spot #3 in this fragment, which has been previously described¹⁵. It is located 76 nucleotides from 5'-end of this fragment.

sequenced the flanking sequences of thirteen random clones of the amplicons. We predominantly found a 22-bp TSD (Fig. 4c), which matched exactly in size and sequence with the TSD found at this insertion hotspot at its genomic location¹⁵. We did not find insertion at any region in the 176-bp fragment other than the hotspot. Therefore, based on three criteria namely, strict requirement of ORF2p expression for mobilization, specific insertion into the retrotransposition hotspot and the 22-bp TSDs accompanying the insertion, we conclude that the events scored by us are due to retrotransposition and not due to DNA recombination. This is the first demonstration of active retrotransposition in a primitive protist.

Analysis of the newly retrotransposed copies of EhSINEs. Next we checked the sequences of the newly retrotransposed copies. To recover all retrotransposition events due to SINEs, we obtained amplicons using the primer pair C1/C2 (Fig. 4d), which gave a 1.4 kb amplicon from the parent plasmid and a 2.0 kb amplicon expected from events where a SINE copy had retrotransposed at the insertion hotspot. We did not obtain bands shorter than 2.0 kb, which shows that in our system the predominant retrotransposition events are contributed by full-length SINEs. This is expected, as truncated SINE transcripts are not seen, whereas full-length transcripts are abundant in *E. histolytica*^{23,24}. This also shows that 5'-truncations

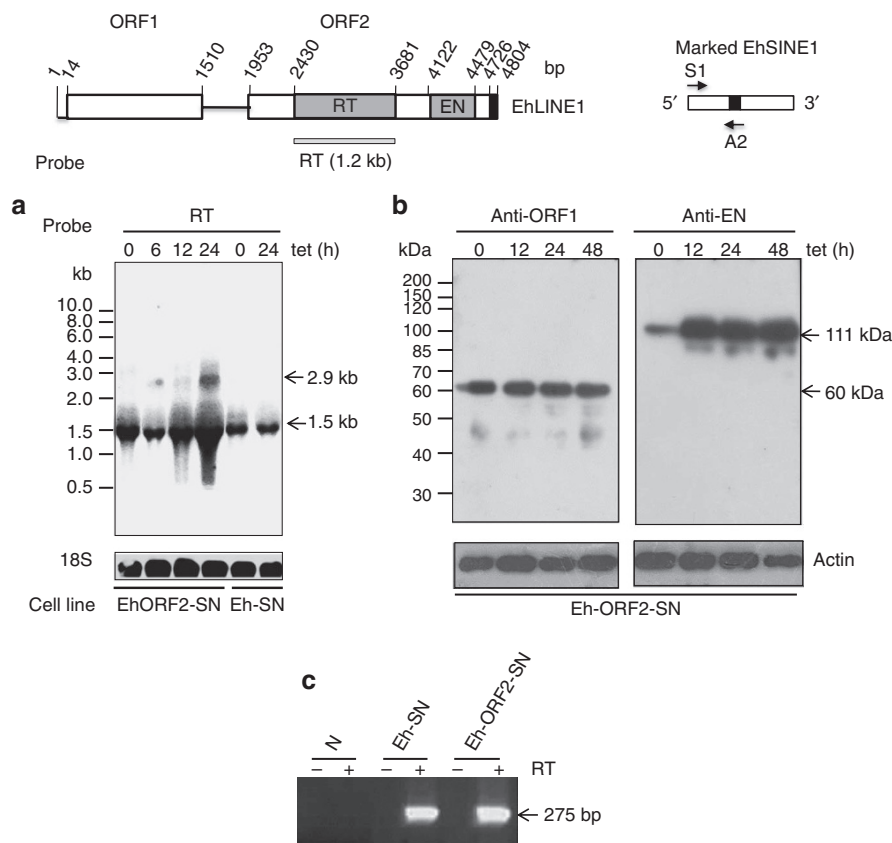


Figure 3 | Expression analysis of EhLINE1 and EhSINE1 in the Eh-ORF2-SN cell line. (a) Northern hybridization was performed with total cellular RNA isolated from cells doubly transfected with pEh-ORF2 and pEh-SN (Eh-ORF2-SN), or with pEh-SN and EhHYG-tetR-O-CAT vector (Eh-SN). Position of the hybridization probe in EhLINE1 is indicated in upper panel. The 2.9-kb band corresponds to the ORF2 transcript after induction with $10 \mu\text{g ml}^{-1}$ tet. No full-length EhLINE1 transcripts of 4.8 kb are seen. (b) Western blotting followed by immunodetection with affinity-purified anti-ORF1 and anti-EN antibodies was performed using total cellular lysate from the Eh-ORF2-SN cell line. Bands of 60 kDa (ORF1p) and 111 kDa (ORF2p) are indicated. Anti-actin antibody was used for loading control. (c) The marked EhSINE1 transcript was detected by RT-PCR with total RNA. RT reaction was with primer A2 from the tag followed by PCR with primers S1 and A2 (position of primers shown in upper panel). The expected size PCR product (275 bp) was obtained. No amplicon was obtained from untransfected cells (N).

are not common during EhSINE retrotransposition. We cut out the 2.0-kb band and reamplified it using the primer pair B1/B2 (Fig. 4). We cloned the 0.8-kb amplicon so obtained, and sequenced 23 randomly selected clones. The data showed that the sequences belonged to three different categories (Table 1) namely; Set I, a set of ten sequences matching completely with the marked SINE; Set II, a set of eight sequences lacking the tag and matching with genomic SINE copies; and Set III, a set of five sequences containing the 25-bp tag at the expected location but, otherwise, matching with genomic SINEs rather than the marked SINE. In these five instances of set III, the tag had associated itself with genomic SINE sequences. Characteristics of each set are as follows. Of the ten sequences in set I, seven were 100% identical to the marked SINE, and three had one mismatch each. The nucleotide position of the mismatch is indicated in Table 1, and shows that the mismatches are not clustered. In set II, all eight sequences lacked the tag and showed only 95–96% sequence identity (19–27 mismatches) with the marked SINE. As this level of identity is seen between random genomic EhSINE1 copies, these sequences would not have arisen from the marked SINE. Rather they may correspond to genomic SINEs. It is estimated that 142 SINE copies are transcribed in *E. histolytica*¹², some of which may be mobilized in our cell line upon tet induction. Sequence comparison showed that seven of the eight sequences in set II indeed showed 98–99% identity with genomic EhSINE1 sequences found

in the *E. histolytica* EST database (3–11 mismatches with best data base hits when compared with the full-length SINE sequence; Table 1, set II). The number of mismatches seen in separate comparisons of 5'-half and 3'-half with the EhSINE1 sequences in the data base is explained later. We confirmed that the database comparison was valid, as random check of PCR amplicons from four of these loci from our cultured cells showed 100% sequence match with the database. These events therefore resulted from the mobilization of transcribed genomic SINE copies and not from the marked SINE copy.

In set III, all five sequences had the 25-bp tag but, surprisingly, showed only 94–95% overall sequence identity (22–27 mismatches) with the marked SINE. When we searched these sequences (minus the tag) for identical hits in the *E. histolytica* data base we found at best 94–98% matches (11–28 mismatches in the full length SINE sequence). However, when we searched the sequence on either side of the tag separately (5'-half and 3'-half of each SINE separately; Table 1, set III), 98–100% matches were obtained, and each side matched with different genomic SINE sequences (accession codes are provided in Supplementary Table S1). Thus it seems that these five sequences are recombinants, derived from at least three different SINE sequences, one of them being the marked SINE and two belonging to different genomic SINEs. If the tag was acquired by the genomic SINE copies through a DNA-recombi-

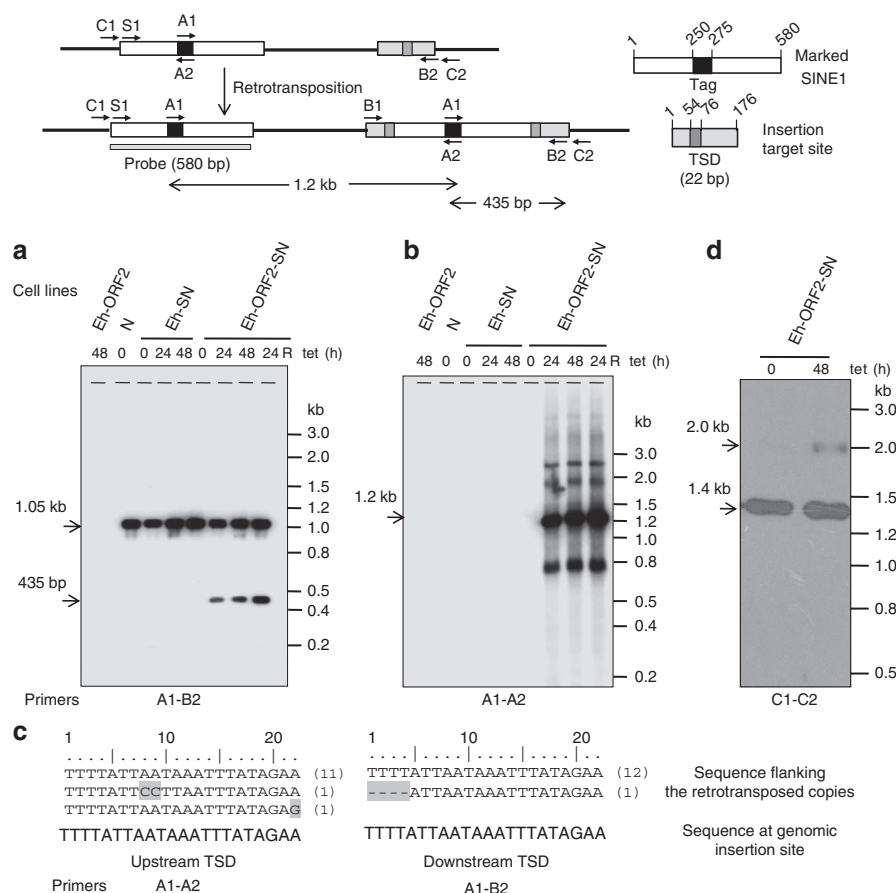


Figure 4 | Tet-induced retrotransposition of the marked EhSINE1 at the target site. The top panel shows the construct containing the marked EhSINE1 (580 bp, hollow rectangle) with a 25-bp GC-rich tag (black rectangle) cloned in the vector pEh-Neo-LUC. The 176-bp fragment (grey box) containing an insertion target site (dark grey box) is shown. Genomic DNA was isolated from untransfected cells (N), single transfectant (Eh-ORF2) and double transfectants (Eh-ORF2-SN and Eh-SN) induced with tet for indicated times. PCR was carried out with the indicated primers whose location is shown in the top panel. PCR products were detected by using double-stranded probe from the marked SINE1. 24R refers to EhORF2-SN cells induced with tet for 24 h and further grown for 24 h in the absence of tet. Results of Southern hybridization are shown in panels **a**, **b** and **d**. Panel **c** shows sequence alignment of upstream and downstream TSDs in 13 retrotransposed copies, obtained by PCR amplification with primers (A1-A2 and A1-B2). The number of copies with the 22-bp TSD identical to the TSD observed at the genomic insertion at the same site is given in brackets. The mismatches are shaded in grey. All retrotransposition events occurred only at indicated target site and nowhere else in the 176-bp fragment.

nation event before retrotransposition, the tag should be present in the transcripts of these SINEs. To check this, we took total RNA from Eh-ORF2-SN cells induced with tet and did RT-PCR in two parts using primers from the tag and from SINE sequences at either end (Fig. 5a,b). We sequenced ten random clones of the amplicons from each side. All clones were identical only to the marked SINE sequence, showing that tag-containing transcripts arose only from the marked SINE copy. To further check if the recombinants existed before the induction of retrotransposition, we performed genomic PCR with DNA from Eh-ORF2-SN cells before tet addition using primers from the tag and opposite primers specific to the five mobilized copies of set III (Fig. 5c). No amplicons were obtained with this DNA, whereas DNA from cells after tet addition gave the expected amplicons, showing that the tag was not associated with these sequences before retrotransposition.

We conclude that recombinant SINEs are formed consequent to retrotransposition. The process is rapid, as we scored these events within 48 h of retrotransposition induction, and occurs at high frequency (>20% of total events scored). Some of the events in set II might also be recombinants, as the number of mismatches reduced

when the 5' and 3' halves of each sequence were searched separately with the database (Table 1). Although the number of mismatches in a full-length comparison ranged from 3 to 20, this number became 0 to 4 when matches were searched only for the 5'-half, and was 0 to 5 when matches were searched for the 3'-half. In two cases the sequence had to be matched in three parts (5', 3', and middle) to get minimum mismatches.

Discussion

Retrotransposition-competent cell lines have greatly assisted in understanding the mechanism of non-LTR retrotransposition *in vivo*^{6–10}. Here we report the first such system in a parasitic protist *E. histolytica*. Our results show that in this organism the newly retrotransposed EhSINE copies undergo high-frequency recombination, not known earlier to take place in this class of non-LTR retrotransposons. Chimeric molecules arising from reverse transcripts have been observed in yeast Ty elements, and were attributed to gene conversion²³, whereas in retroviruses high-frequency recombination occurs during reverse transcription of the two co-packaged RNAs in the virion as a result of template switching²⁴. In non-LTR elements, tripartite chimeric LINES have earlier been reported in

Table 1 | Sequence analysis of newly retrotransposed SINE copies.

		Number of mismatches with the best database hits [†]							
		Full length		5'-end		Middle		3'-end	
Sequence number	Number of mismatches with marked SINE (position)	Length matched (bp)	Mis-matches	Length matched (bp)	Mis-matches	Length matched (bp)	Mis-matches	Length matched (bp)	Mis-matches
Set I: Marked SINE1									
1-7	0	NA	NA	NA	NA	NA	NA	NA	NA
8	1 (134)	NA	NA	NA	NA	NA	NA	NA	NA
9	1 (274)	NA	NA	NA	NA	NA	NA	NA	NA
10	1 (482)	NA	NA	NA	NA	NA	NA	NA	NA
		Number of mismatches with the best database hits [†]							
		Full length		5'-end		Middle		3'-end	
Sequence number	Number of mismatches with marked SINE	Length matched (bp) [‡]	Mis-matches	Length matched (bp) [‡]	Mis-matches	Length matched (bp) [‡]	Mis-matches	Length matched (bp) [‡]	Mis-matches
Set II: Genomic SINEs									
1	22	501	11	340	4	NA	NA	161	0
2	25	503	6	370	4	NA	NA	133	0
3	19	502	5	305	4	NA	NA	197	1
4	24	503	3	325	0	NA	NA	178	1
5	27	502	20	125	2	255	4	122	5
6	24	502	11	350	2	70	3	82	1
7	22	502	10	380	3	NA	NA	122	2
8	19	502	5	214	1	NA	NA	288	1
		Number of mismatches with the best database hits [†]							
		Full length		5'-end		Middle		3'-end	
Sequence number	Number of mismatches with marked SINE	Length matched (bp) ^{‡,§}	Mis-matches	Length matched (bp) [‡]	Mis-matches	Length matched (bp)	Mis-matches	Length matched (bp) [‡]	Mis-matches
Set III: Recombinant SINEs									
1	22	476	16	196	0	NA	NA	280	5
2	22	476	15	196	0	NA	NA	280	1
3	25	476	18	196	0	NA	NA	280	4
4	24	502	11	222	2	NA	NA	280	4
5	27	450	28	170	1	NA	NA	280	0

NA, not applicable.

The E. histolytica data base was searched for matches with the retrotransposed copies. Comparison was made either for the full-length SINE sequence on in parts (5'-end, middle, 3'-end) as indicated.

[†] See Supplementary Table S1 for accession code of each hit.

[‡] 25 nucleotides were removed from both ends for BLAST analysis.

[§]Length variation is due to different numbers of internal 26-mer repeats in the EhSINE copies¹².

NA, not applicable.

The *E. histolytica* data base was searched for matches with the retrotransposed copies. Comparison was made either for the full-length SINE sequence or in parts (5'-end, middle, 3'-end) as indicated.[†] See Supplementary Table S1 for accession code of each hit.[‡] 25 nucleotides were removed from both ends for BLAST analysis.[§] Length variation is due to different numbers of internal 26-mer repeats in the EhSINE copies¹².

a fungal genome²⁵, and in mammalian genomes U6/L1 pseudogene chimeras have been experimentally demonstrated²⁶. However, recombination between multiple copies of the same SINE family during retrotransposition is a novel observation.

The demonstrated properties of RT to displace the RNA template during complementary DNA (cDNA) synthesis, and to perform multiple template jumping^{27,28} could lead to these recombinants. The template jumping activity reported for non-LTR retrotransposons involves end-to-end jumping between the template RNAs, and this is frequently accompanied by the addition of non-templated nucleotides at the junction. The recombinant SINEs reported here are not generated by end-to-end jumping. They would require the RT to switch at internal positions from one template RNA to the next. Such switching has not been reported for the RT encoded by non-LTR elements and needs to be explored further. Our data (Table 1) suggests that, more frequently, no switches take place during retrotransposition (at least 10 out of 23 occurrences; set I); but multiple switching events are commonly encountered. In set II, four events (sequences 2, 3, 4 and 8) probably

did not involve any switch, whereas two (numbers 1 and 7) could involve one switch each, and the remaining two (numbers 5 and 6) could involve two switches each. In set III, all five events seemed to be a result of two switches. This may be due to the small sample size in this set.

If such recombination is indeed common, one should expect the *E. histolytica* SINE population to display a prominent mosaic structure. To check this, we did multiple alignment of 63 full-length EhSINE1 copies thought to have retrotransposed most recently¹², and looked at all positions that were polymorphic in at least 20% of the SINEs. We found 16 such positions. We organized the sequences in these positions in blocks of four, and clustered the positions with identical sequences into sets in the leftmost block (Fig. 6). Upon aligning these sequences with the next block of four, it is evident that the sequences of identical sets in the first block segregate out and associate in various combinations. Nine sets in the first block associated with twelve sets in the second block in 34 different combinations. The combination of patterns in all four blocks taken together showed clear indication of mosaic formations that would

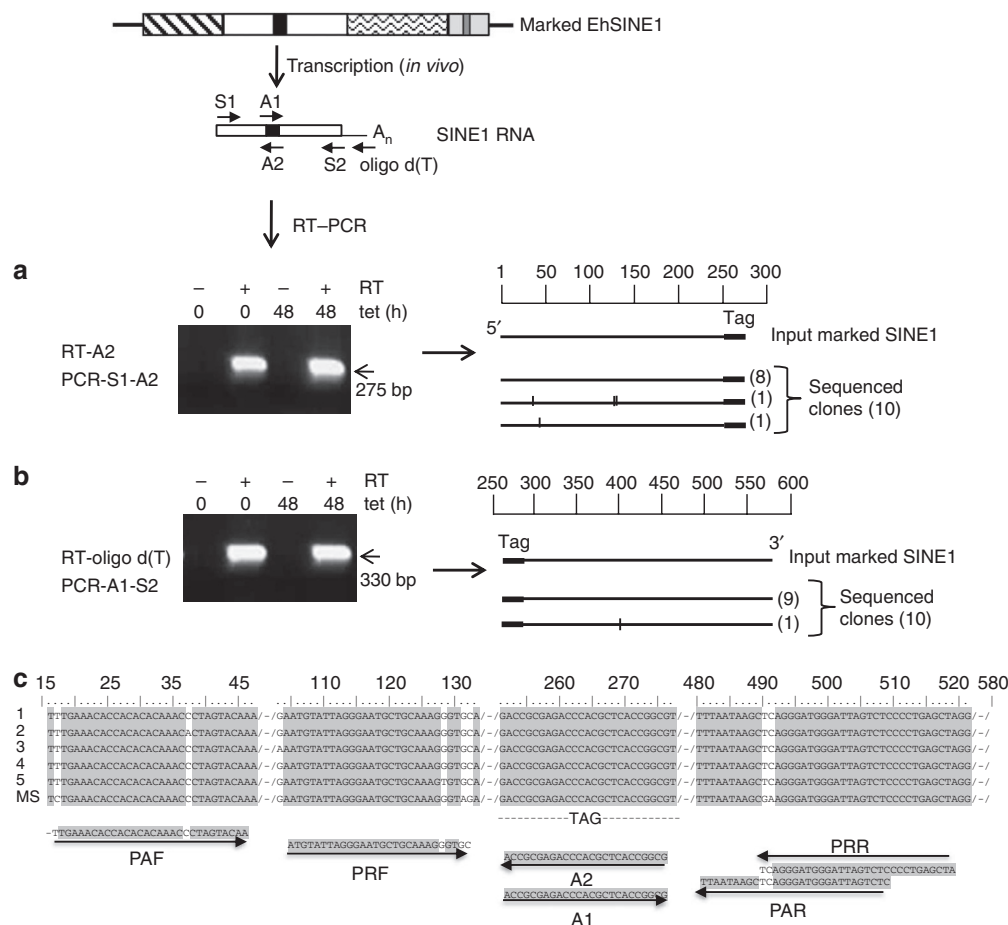


Figure 5 | Sequence changes in SINEs were not due to transcription or genomic recombination. Total cellular RNA was isolated from the Eh-ORF2-SN cells grown for 48 h with $10 \mu\text{g ml}^{-1}$ tet and cDNA was synthesized with MuMLV-RT by using primers shown in the upper panel. Primers S1 and S2 are from conserved SINE sequences and would amplify genomic as well as marked SINE1. **(a)** RT reaction with primer A2 followed by PCR with primers S1-A2. The amplicon of 275 bp thus obtained was cloned in pGEMTEasy vector and ten random clones were selected for sequencing. The DNA sequence of SINE1 in these clones is represented by bold lines with eight clones having sequence identical to input-tagged EhSINE1 and two clones having three and one mismatches, respectively (represented by vertical lines). **(b)** Same as in **(a)**, with indicated RT-PCR primers to obtain the SINE sequence 3' of the tag. As EhSINEs are polyadenylated at 3' ends^{21,22}, RT reaction was primed with oligo d(T). **(c)** Multiple alignment of the five clones of set III and the marked EhSINE1 (MS) shows the location of primers for genomic PCR. Primer pairs PRF-A2 and A1-PRR were used with total genomic DNA to check if tag-containing set III sequences already existed in the genome before induction of retrotransposition with tet. Primer pair PAF-PAR was derived from conserved sequences and served as positive control.

be expected to arise from internal template jumps during retrotransposition.

Although sequence analysis of human Alu subfamilies showed the existence of mosaic elements²⁹, these were not experimentally found⁷⁻⁹. Mosaics could have been missed in these studies because of the smaller number of retrotransposition events scored, and the selection pressure used. Our results are the first direct demonstration that SINE copies engage in active sequence exchange during retrotransposition, leading to the rapid spread of the sequence tag to the SINE population, and generation of diversity. As mRNA transcripts are also templates of the same retrotransposition machinery during retropseudogene formation, it will be interesting to see if mRNA transcripts could also engage in similar recombination during reverse transcription²⁹.

Methods

Cell culture and growth conditions. Trophozoites of *E. histolytica* strain HM-1: IMSS (clone 6) were axenically maintained and stable transfectants were obtained

as described previously³⁰. Single and double transfectants were maintained with 10 µg ml⁻¹ of Hygromycin B or G418 or both.

Immunodetection. Western analysis was performed with 100 µg of total cell lysates separated on 10% SDS–polyacrylamide gel electrophoresis and blotted using Mini-Trans Blot Electrophoretic Cell (Bio-Rad). Polyclonal antibodies, anti-EN (mouse) and anti-ORF1 (rabbit) were raised against each his-tagged recombinant protein purified through Ni-NTA agarose. 1:5,000 dilution of each anti-serum followed by HRP-conjugated secondary antibodies (1:10,000, Sigma) were used to detect respective polypeptides. ECL reagents were used for visualization (Millipore).

DNA sequence analysis. The genome sequence of *E. histolytica* having 1529 scaffolds was downloaded from NCBI database, with accession IDs AAFB000000000, and was searched (using BLAST tool) for matches with the sequences of retrotransposed copies of Sets II and III (Table 1 and Supplementary Table S1). In all, 25 nucleotides from both the ends of these copies were removed before BLAST analysis to remove any bias due to other factors (for example, addition of non-templated nucleotides during the RT reaction). Hits showing 100% coverage and maximum similarity were recorded. The coordinates of EhsINES in various scaffolds are as per previous report¹².

Sr. no.	Scaffold	Coordinates	Positions													
			57	96	163	223	238	313	403	404	414	434	439	444	445	453
1	DS571451	3493–2943	A	C	A	C	G	G	G	C	G	A	A	G	A	T
2	DS572264	932–387	A	C	A	G	A	T	G	C	C	C	C	T	A	C
3	DS571279	21678–21129	A	C	A	G	A	T	G	C	C	A	A	G	A	T
4	DS572093	1734–1187	A	C	A	G	A	T	G	C	G	C	C	T	G	C
5	DS571347	7763–8309	A	C	A	G	G	T	G	G	C	A	A	G	A	T
6	DS571347	2150–2696	A	C	A	G	G	T	G	G	C	A	A	G	A	T
7	DS571487	12336–12883	A	C	A	G	A	T	G	C	G	C	C	T	G	C
8	DS571145	332477–331932	A	C	A	T	G	T	G	C	C	A	A	G	A	T
9	DS571147	112278–111732	A	C	A	T	G	T	G	C	C	C	C	T	G	C
10	DS571221	45970–46517	A	C	A	T	G	T	A	C	C	T	C	G	A	T
11	DS571233	53658–54208	A	C	A	T	A	T	G	C	C	A	A	G	A	T
12	DS571643	734–1284	A	C	A	T	G	T	G	C	G	T	C	T	G	A
13	DS571167	34094–34639	A	C	A	T	A	G	G	C	C	C	C	T	G	A
14	DS571668	4836–4290	A	C	A	T	G	T	A	C	C	C	C	T	G	C
15	DS571192	24324–24871	A	C	A	T	G	G	A	C	C	A	A	G	A	T
16	DS571151	89505–88960	A	C	G	G	G	G	G	C	C	C	C	T	G	A
17	DS571483	7440–6892	A	C	G	T	A	T	A	C	C	C	C	T	G	A
18	DS571266	27732–28282	A	C	G	T	G	G	A	C	C	A	A	G	A	T
19	DS571832	:655–1201	A	C	G	T	G	T	G	C	C	C	C	G	A	C
20	DS571372	19793–19246	A	C	G	T	G	T	G	G	C	A	A	G	A	T
21	DS571530	1520–2063	A	C	G	T	G	T	A	G	C	C	C	G	A	T
22	DS571240	35235–34688	A	T	A	T	G	G	G	C	G	A	A	G	A	T
23	DS571255	5733–5183	A	T	A	T	A	T	G	C	C	C	C	T	G	A
24	DS571284	36571–37118	A	T	A	T	A	T	G	C	G	C	C	T	G	A
25	DS571249	2300–2846	A	T	A	T	G	G	G	G	G	A	A	T	G	T
26	DS571166	33449–33996	A	T	A	T	G	T	G	C	C	A	A	G	A	T
27	DS571489	6580–7122	A	T	A	T	G	T	G	C	A	C	C	T	G	A
28	DS571146	102828–103375	A	T	A	T	A	T	G	C	C	T	C	T	G	A
29	DS571158	27154–27704	A	T	A	T	G	T	G	C	C	A	A	G	A	T
30	DS571240	22360–21813	A	T	A	T	G	T	G	C	G	A	A	G	A	T
31	DS571257	8157–7609	A	T	A	T	G	T	G	C	G	A	A	G	A	T
32	DS571237	34661–35208	A	T	A	T	G	G	A	C	C	C	C	T	G	A
33	DS571219	15122–14577	G	A	A	A	A	T	G	T	C	A	T	A	G	A
34	DS571226	27023–26476	T	A	A	T	G	T	G	C	C	A	C	T	G	A
35	DS571227	51623–51076	T	C	A	T	G	T	G	C	G	A	C	T	G	A
36	DS571410	9760–9214	T	C	A	T	G	T	G	C	G	A	A	T	G	T
37	DS571445	9720–9173	T	C	A	T	G	T	G	G	C	A	A	T	G	T
38	DS571508	4500–3954	T	C	A	T	G	T	A	C	C	C	C	T	G	A
39	DS571272	1423–1970	T	C	A	T	A	T	G	G	C	A	A	G	A	T
40	DS571358	20913–21457	T	C	A	T	A	T	G	C	C	A	A	G	A	C
41	DS571145	96823–97370	T	C	A	T	A	T	A	C	C	A	A	G	A	T
42	DS571164	127996–128542	T	C	A	T	G	G	G	C	C	A	A	G	G	A
43	DS571179	37873–38420	T	C	A	T	G	T	G	G	C	A	A	G	A	T
44	DS571190	67757–68303	T	C	A	T	G	T	G	C	C	C	C	T	G	C
45	DS571247	15177–14629	T	C	A	T	G	G	G	C	G	A	A	G	A	T
46	DS571256	33189–32642	T	C	G	T	G	T	G	G	C	C	C	T	G	C
47	DS571382	18914–18367	T	C	G	T	G	T	A	C	C	A	A	G	A	T
48	DS571494	2811–3357	T	C	G	T	G	T	G	C	C	C	C	T	G	C
49	DS571228	4744–5290	T	T	A	C	A	T	A	G	C	C	C	T	G	C
50	DS571277	39116–38571	T	T	A	G	A	T	G	C	C	C	C	T	G	A
51	DS572252	2–548	T	T	A	G	A	T	G	C	C	C	C	T	G	A
52	DS572593	697–150	T	T	A	T	G	T	G	C	C	A	A	G	A	T
53	DS571423	1030–483	T	T	A	T	G	T	G	C	C	A	A	G	A	T
54	DS571278	6338–5791	T	T	A	T	G	T	A	G	C	A	A	G	A	T
55	DS571148	58191–58738	T	T	A	T	G	T	A	G	C	C	C	T	G	C
56	DS571317	27553–28099	T	T	A	T	G	T	A	C	C	A	C	T	G	A
57	DS571310	11751–11204	T	T	A	T	A	T	G	C	C	C	C	T	G	C
58	DS571214	15165–14617	T	T	A	T	A	G	A	C	C	A	C	G	A	C
59	DS571257	45861–46408	T	T	A	T	A	T	G	C	C	C	C	T	G	A
60	DS571346	26099–25553	T	T	A	T	G	T	A	C	C	A	A	G	A	T
61	DS571186	24988–25538	T	T	G	C	G	G	G	C	C	A	A	G	A	T
62	DS571151	154899–154352	T	T	G	T	G	T	G	C	C	A	A	T	G	T
63	DS571183	68021–67471	T	T	G	T	A	G	G	G	C	T	A	G	A	T

Figure 6 | Polymorphic sites in the EhSINE1 population are distributed in a mosaic pattern. The coordinates of the 63 full-length EhSINE1 copies selected for analysis are shown on the left. These copies were selected because they are thought to have retrotransposed most recently¹². Sequences of the 63 copies were aligned and nucleotide positions varying in at least 12 copies were selected. In all, 16 such positions found are displayed. Blocks of four consecutive positions with the same sequence have been given the same colour for clarity. The five identical pairs found are indicated by same shading in the left column.

References

- Singer, M. F. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28**, 433–434 (1982).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Cordaux, R., Hedges, D. J. & Batzer, M. A. Retrotransposition of Alu elements: how many sources? *Trends Genet.* **20**, 464–467 (2004).
- Roy, A. M. *et al.* Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**, 1485–1495 (2000).
- Salem, A. H., Ray, D. A., Hedges, D. J., Jurka, J. & Batzer, M. A. Analysis of the human Alu Ye lineage. *BMC Evol. Biol.* **5**, 18 (2005).
- Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
- Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).
- Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400 (2011).
- Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).
- Kajikawa, M. & Okada, N. LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111**, 433–444 (2002).
- Bakre, A. A., Rawal, K., Ramaswamy, R., Bhattacharya, A. & Bhattacharya, S. The LINEs and SINEs of *Entamoeba histolytica*: comparative analysis and genomic distribution. *Exp. Parasitol.* **110**, 207–213 (2005).
- Huntley, D. M., Pandis, I., Butcher, S. A. & Ackers, J. P. Bioinformatic analysis of *Entamoeba histolytica* SINE1 elements. *BMC Genomics* **11**, 321 (2010).
- Lorenzi, H. *et al.* Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. *BMC Genomics* **9**, 595 (2008).
- Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16**, 793–805 (1999).
- Mandal, P. K., Bagchi, A., Bhattacharya, A. & Bhattacharya, S. 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 (2004).
- Van Dellen, K., Field, J., Wang, Z., Loftus, B. & Samuelson, J. LINEs and SINE-like elements of the protist *Entamoeba histolytica*. *Gene* **297**, 229–239 (2002).
- Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
- Martin, S. L., Li, J., Epperson, L. E. & Lieberman, B. Functional reverse transcriptases encoded by A-type mouse LINE-1: defining the minimal domain by deletion analysis. *Gene* **215**, 69–75 (1998).
- Yang, J., Malik, H. S. & Eickbush, T. H. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc. Natl Acad. Sci. USA* **96**, 7847–7852 (1999).
- Moran, J. V. & Gilbert, N. Mammalian LINE-1 retrotransposons and related elements. In *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 836–869 (American Society for Microbiology, 2002).
- Cruz-Reyes, J., ur-Rehman, T., Spice, W. M. & Ackers, J. P. A novel transcribed repeat element from *Entamoeba histolytica*. *Gene* **166**, 183–184 (1995).
- Willhoeft, U., Buss, H. & Tannich, E. The abundant polyadenylated transcript 2 DNA sequence of the pathogenic protozoan parasite *Entamoeba histolytica* represents a nonautonomous non-long-terminal-repeat retrotransposon-like element which is absent in the closely related nonpathogenic species *Entamoeba dispar*. *Infect. Immun.* **70**, 6798–6804 (2002).
- Derr, L. K. & Strathern, J. N. A role for reverse transcripts in gene conversion. *Nature* **361**, 170–173 (1993).
- Delviks-Frankenberry, K. *et al.* Mechanisms and factors that influence high frequency retroviral recombination. *Viruses* **3**, 1650–1680 (2011).
- Gogvadze, E., Barbisan, C., Lebrun, M. H. & Buzdin, A. Tripartite chimeric pseudogene from the genome of rice blast fungus *Magnaporthe grisea* suggests double template jumps during long interspersed nuclear element (LINE) reverse transcription. *BMC Genomics* **8**, 360 (2007).
- Garcia-Perez, J. L., Doucet, A. J., Bucheton, A., Moran, J. V. & Gilbert, N. Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res.* **17**, 602–611 (2007).
- Bibillo, A. & Eickbush, T. H. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* **316**, 459–473 (2002).
- Bibillo, A. & Eickbush, T. H. End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J. Biol. Chem.* **279**, 14945–14953 (2004).
- Carroll, M. L. *et al.* Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.* **311**, 17–40 (2001).
- Hamann, L., Buss, H. & Tannich, E. Tetracycline-controlled gene expression in *Entamoeba histolytica*. *Mol. Biochem. Parasitol.* **84**, 83–91 (1997).

Acknowledgements

We acknowledge Jitender Kumar for helping to raise the anti-ORF1 antibody. V.P.Y. is a recipient of a Senior Research Fellowship from Council of Scientific and Industrial Research, Government of India. This work was funded by grant to S.B. from Department of Science and Technology and Department of Biotechnology, India.

Author contributions

S.B. supported and designed the research and wrote the manuscript along with V.P.Y.; A.B. contributed in designing the research and in the computational analysis; V.P.Y. designed and performed the experiments; P.K.M. carried out the reconstruction of ORF2 and prepared the anti-EN antibodies. All authors have read and approved the manuscript.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Yadav, V. P. *et al.* Recombinant SINEs are formed at high frequency during induced retrotransposition *in vivo*. *Nat. Commun.* **3**:854 doi: 10.1038/ncomms1855 (2012).