Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 333 (2005) 230–240
BBRC
www.elsevier.com/locate/ybbrc

Identification of 18 new transcribed retrotransposons in *Schistosoma mansoni*

Ricardo DeMarco [a], Abimael A. Machado [b], Alexandre W. Bisson-Filho [a], Sergio Verjovski-Almeida [a,b,*]

[a] *Departamento de Bioquimica, Instituto de Quimica, Universidade de São Paulo, 05508-900 São Paulo, SP, Brazil*
[b] *Laboratorio de Bioinformatica, Departamento de Bioquimica, Instituto de Quimica, Universidade de São Paulo, 05508-900 São Paulo, SP, Brazil*

Received 11 May 2005
Available online 31 May 2005

Abstract

This work describes 18 new transcribed retrotransposons of the blood fluke *Schistosoma mansoni*. Among them, 9 were LTR, 8 non-LTR, and 1 *Penelope*-like element (PLE) retrotransposon. Sequences were generated by in silico reconstruction using *S. mansoni* ESTs and transcripts obtained by rapid amplification of cDNA ends, complemented in some cases by sequencing of genomic clones amplified by PCR. A novel element from the ancient R2/R4/CRE transposon group is described for the first time in *S. mansoni*. In addition, one non-LTR retrotransposon family displays long (40–450 bp) 3′-UTR with at least six different transcribed sequences among the copies, five LTR retrotransposons have abundantly transcribed incomplete copies lacking the sequence segment coding for the reverse transcriptase domain, and four non-LTR retrotransposons code for DNA-binding PHD domains that may give them a differential targeting. These results allow for a comprehensive description of the transcribed retrotransposon diversity of this complex human parasite.

Keywords: *Schistosoma mansoni*; Phylogenetic analyses; LTR retrotransposon; Non-LTR retrotransposon; R2 family retrotransposon; Transcribed retrotransposons

Schistosoma mansoni, a digenetic platyhelminth trematode, is the primary causative agent of schistosomiasis in humans, being an important factor of morbidity in the world. The disease caused by this blood fluke is endemic in 74 developing countries infecting about 200 million individuals, and it is estimated that an additional 500–600 million are at risk [1]. The genome of *Schistosoma* has approximately 270 Mbp [2] and a considerable portion (more than 20%) is estimated to be composed of retrotransposons [3].

Ten *S. mansoni* retrotransposons have been characterized. Six of them are LTR retrotransposons, also known simply as retrotransposons [4], of which four are from the Gypsy/Ty3 family and two from the Bel family [5–8]. Other three elements are non-LTR retrotransposons, also known as retroposons [4] of the CR1 and the RTE1 families [5,9,10]. The remaining tenth element, Cercyon, has been briefly described [11] and belongs to the recently characterized *Penelope*-like elements (PLE) class of retrotransposons, which were shown to retain introns and present some structural peculiarities [11]. In addition, a classical DNA transposon of the Merlin family has recently been detected in the *S. mansoni* genome, and several lines of evidence of its activity were shown [12]. Four of these elements have been described as having low genomic copy number and high transcriptional activity, suggesting that active copies of such retrotransposons may still

* Corresponding author. Fax: +55 11 3091 2186.
E-mail address: verjo@iq.usp.br (S. Verjovski-Almeida).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.05.080

exist in the genome [5]. Although no transposition event has been effectively demonstrated for any of these retrotransposons, the presence of reverse transcriptase activity in *Schistosoma* extracts suggests the existence of elements with intact protein domains [13].

Although initially regarded as a selfish DNA with negative impact on the host [14,15], transposable elements are one of the principal forces driving the evolution of eukaryotic genomes [14] and the generation of phenotypic diversity [16,17], their significant contribution to gene evolution being demonstrated [18,19]. Retrotransposons have also been demonstrated to facilitate DNA repair [20,21] and serve as substitutes for telomeres [22].

The use of fragments of several copies of a transposable element to obtain an in silico consensus sequence is a commonly employed approach [23]. The reconstruction process aims at obtaining a consensus sequence lacking insertions, deletions, and stop codons that are otherwise accumulated in the inactive copies. In fact, a copy of *Sleeping Beauty*, a Tc1-like transposon, was physically reconstructed from its inactive copies and demonstrated to be transpositionally active in vitro [24]. The use of EST sequences to reconstruct transposable elements permits a direct access to their transcriptionally active copies, which are less prone to contain truncations and substitutions common to most of the genomic copies, thus facilitating the reconstruction process [5].

In this work, we describe 18 new transcribed retrotransposons, thus significantly increasing the knowledge about retrotransposable elements transcribed in *S. mansoni*. The expressed *S. mansoni* retrotransposons were divided into six major groups (i.e., having more than one retrotransposon per group) and five minor groups. Identification of distinctive traits in some of the retrotransposons suggested that different solutions were prevalent during the adaptation of these elements along *S. mansoni* evolution.

Materials and methods

Reconstruction of retrotransposon sequences. Retrotransposons were reconstructed from *S. mansoni* retrotransposon ESTs as described [5], with the exception that we also used reads of BAC ends deposited in GenBank only to obtain information for short internal regions of some retrotransposons, as described under Results. The 6599 *S. mansoni* retrotransposon ESTs used as a starting point in the reconstruction shown in the present work were deposited in GenBank under Accession Nos. CF497203–CF503801. Reconstructed sequences for the following new retrotransposons are available in the supplementary methods and also at our website (http://verjo2.iq.usp.br/): LTR retrotransposons Saci-4, Saci-5, Saci-6, Saci-7, Nonaut-1, Nonaut-3, Nonaut-4, Nonaut-5, Nonaut-5.1, Nonaut-6.1, Nonaut-6.2; non-LTR retrotransposons Perere-2, Perere-3, Perere-4, Perere-5, Perere-6, Perere-7, Perere-8, Perere-9; PLE retrotransposon Perere-10.

Construction of phylogenetic trees. Alignment of the reverse transcriptase domain of new and known *S. mansoni* retrotransposons was obtained with Clustal X program (v 1.83). Further analysis with Clustal X using the neighbor-joining method excluding positions with gaps resulted in the phylogenetic trees shown in the Figures. Phylogenetic trees were drawn using Tree View program (v. 1.6.6). The GenBank sequences utilized for construction of alignments and phylogenetic trees are given in the supplementary methods.

Amplification of genomic clones. Genomic clones for Perere-9 and -10 were obtained using genomic DNA as template in a PCR using specific primers designed from reconstructed sequences of retrotransposons. These primers were used for PCR of *S. mansoni* genomic DNA using Advantage 2 polymerase mix (BD Bioscience) and the following cycling program: 95 °C for 3 min; 40 cycles of 95 °C for 30 s, 60 °C for 30 s, and 68 °C for 5 min in a GenAmp PCR system 9700 (Applied Biosystems). The ramp of temperature transition between annealing and extension steps was reduced to 5% of its default speed in order to increase the amount of amplified product. The products were analyzed in 1.2% agarose gel and cloned in pGEM-T vector (Promega) for further sequencing. Genomic sequences for Perere-9 and -10 were deposited in GenBank under Accession Nos. AY838781 and AY838774, respectively.

Rapid amplification of cDNA ends. mRNA was obtained from adult parasites conserved in RNAlater (Ambion) by extraction of tissue with MACs mRNA isolation kits (Miltenyi Biotech). Two hundred nanograms of mRNA was used for reverse transcription using the protocol of the 3′RACE system kit for rapid amplification of cDNA ends (Invitrogen) and specific primers based on the reconstructed retrotransposon sequence. PCR was performed with Advantage II (BD Biosciences) with the buffer supplied by the manufacturer, 200 μM dNTPs and 200 nM of each primer using the following cycling: 95 °C for 1 min plus 35 cycles each at 95 °C for 30 s, 55 °C for 30 s, and 68 °C for 3 min, followed by a final extension at 68 °C for 3 min. The products were analyzed in 1.2% agarose gel and cloned in pGEM-T vector (Promega) for further sequencing. Partial sequences for clones generated by RACE were deposited in GenBank under Accession Nos. Perere-3, AY838775 to AY838780 (six different clones); Nonaut-5, AY838773.

Results and discussion

Schistosoma mansoni retrotransposon diversity

As part of the effort to characterize the *S. mansoni* transcriptome, approximately 160,000 *S. mansoni* ESTs were generated, putative retrotransposon EST sequences were filtered out, and a total of 120,000 ESTs were analyzed [25]. The database of putative retrotransposon EST sequences had not been assembled and analyzed in our original work, due to the complexity arising from the high degree of variability within the transposon families. Subsequently, 4398 ESTs contained in the database of putative transposon EST sequences were readily assembled into four different retrotransposons, characterizing them as highly transcribed elements (Saci-1, -2, and -3, and Perere) [5]. A total of 6599 additional ESTs were identified as retrotransposon sequences, but they proved to be highly diverse and fragmentary. These data suggested the presence of a large number of different transcribed elements and prompted us to perform additional experiments, as now described, to characterize their full-length sequences.

The coding region of 18 new transcribed *S. mansoni* retrotransposons was reconstructed, as described in detail in the following sections. Blastn comparison of all 18 new *S. mansoni* retrotransposon sequences to each other and to the 10 previously known retrotransposons showed very little or no similarity at the DNA level (87% maximum similarity over a stretch representing a maximum of 1.1% of their full-length), justifying the description of a total of 28 distinct *S. mansoni* retrotransposon elements. In fact, a phylogenetic analysis of 24 of these *S. mansoni* retrotransposons, which code for the Reverse Transcriptase domain, showed that they are all distantly related and clustered together with a number of representative retrotransposons from different species (Fig. 1). Among the new elements, five of them were identified as new members of LTR retro-

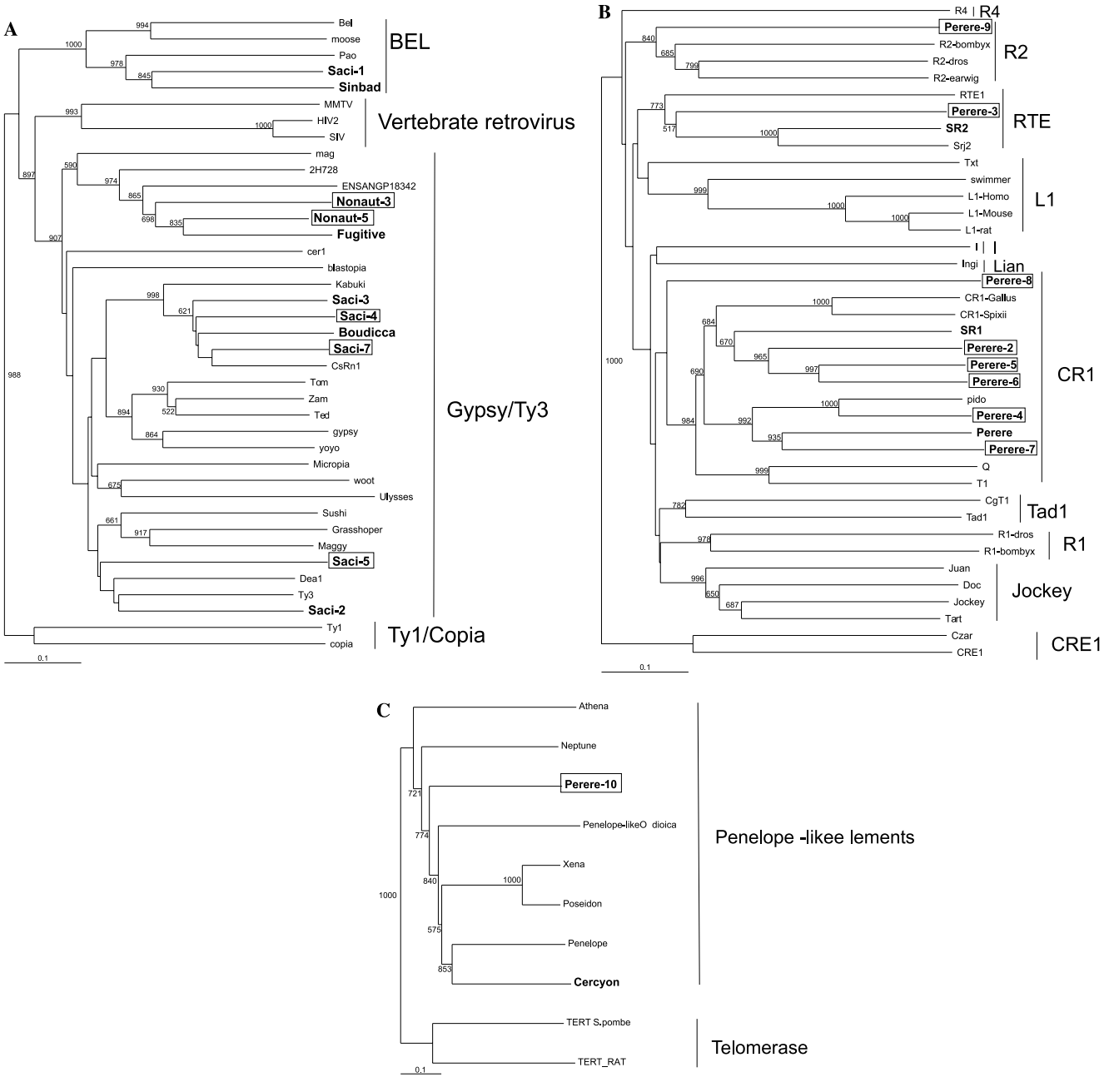

Fig. 1. Phylogenetic trees for the reverse transcriptase domains of *S. mansoni* retrotransposons. (A) LTR-retrotransposons; (B) non-LTR retrotransposons; (C) *Penelope*-like element (PLE) retrotransposons. All 24 *S. mansoni* retrotransposons are in bold characters; the 14 new *S. mansoni* retrotransposons identified in this work that have the reverse transcriptase domain are marked with boxes. Four new *S. mansoni* retrotransposons do not have the reverse transcriptase domain and could not be included in this analysis. Trees were constructed by the neighbor-joining method. Numbers represent the confidence of the branches assigned by bootstrap analysis (in 1000 samplings); bootstrap values lower than 500 are omitted from the figure.

transposon families known to be present in *S. mansoni* (Fig. 1A), eight were of the non-LTR group (Fig. 1B), one of them being a member of the R2 family, a novel family in *S. mansoni*, and one was a *Penelope*-like element (PLE) retrotransposon (Fig. 1C).

BLASTN comparison against the preliminary genome sequence assembly from the Sanger Institute (January 14, 2005 release) showed that the best genome sequence hit for each of the 18 new retrotransposons had between 93% and 99% identity with the corresponding retrotransposon reconstructed sequence obtained here. The genomic DNA sequence covered from 82% to 100% of the full-length of each of the 18 new retrotransposons, with the exception of Saci-6, which had a genomic sequence coverage of only 68%. These results confirm the overall structure of the retrotransposon sequences reconstructed here from EST transcripts; the lack of an exact genomic matching and the partial coverage obtained may be explained by the partial nature of the present genomic assembly, in which the genomic sequence is still fragmented into 70,714 contigs (January 14, 2005 release; ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/). In some cases, it may also represent sequence errors in the reconstructed copies due to the single-pass nature of EST sequencing.

Diversity of LTR retrotransposons

Four out of the nine new LTR retrotransposons identified here (Saci-4, -5, and -7, and Nonaut-5) coded for all protein domains characteristic of LTR retrotransposons, namely reverse transcriptase, protease, RNase H, gag, and integrase (Fig. 2A), and the detailed sequence alignments can be seen in Supplementary Fig. 1. An additional LTR retrotransposon, Nonaut-3, was reconstructed with a partial ORF lacking the stop codon at the 3′-end; the deduced sequence has an incomplete RNase H domain and lacks the integrase domain supposed to be present at the carboxyl-terminal end (Fig. 2B). However, we were able to identify in the *S. mansoni* preliminary genome assembly obtained at the Sanger Institute one sequence (333748.c003209129.Contig2) with 100% coverage and 99% identity to our partially reconstructed sequence, which in addition extends the reconstructed sequence and reveals an intact integrase domain. Therefore, although we obtained a partial sequence we will refer to Nonaut-3 as a complete element.

The remaining four LTR retrotransposons (Saci-6, Nonaut-1.1, -4.1, and -6.1/-6.2) each has an integral ORF that is shorter than usual, and the deduced proteins lack the reverse transcriptase domain (Fig. 2C). Therefore, they were not included in the phylogeny analysis, as mentioned before, and were identified as LTR elements by the overall structure and by the presence of terminal repeats (Fig. 2C). Because the reconstructed copies included a defined stop codon and the LTR element at both ends, we assume that these sequences represent fully reconstructed transcripts of truncated elements. If complete non-truncated copies exist in the genome, they seem not to be transcribed at a significant level. Further studies are warranted to determine whether these truncated elements are capable of transposing by use in *trans* of the proteins of intact copies or of other elements.

Nonaut, a group of LTR retrotransposons with abundantly transcribed truncated copies with an intact gag domain

Two elements within the Ty-3/Gypsy family (Nonaut-3, -5) together with the recently described Fugitive retrotransposon [6] represent a new LTR group in *S. mansoni*, as evidenced by the phylogeny analysis (Fig. 1A). Three other Nonaut truncated members probably belong to this group (Nonaut-1.1, -4.1, and -6.1/-6.2), as seen by the conserved gag and protease domains (Supplementary Fig. 1). This LTR group is different from the two groups previously known in the Ty-3/Gypsy family in *S. mansoni*, one that is formed by Bouddica/Saci-3 and the other containing Saci-2 (Fig. 1A).

Among the complete and the shorter truncated copies of Nonaut elements, the only deduced domain present throughout all members is a conserved segment coding for a gag protein (Figs. 2A–C). This conservation suggests that the gag protein is being produced in excess in relation to the other diverse components of pol. It is known that for virus assembly and production of virion particles gag is expressed in excess in relation to pol protein [26,27] and a mechanism for generating excess of gag protein in relation to integrase was also described for Tf1 retrotransposon [28]. It is tempting to hypothesize that also in *S. mansoni* a high expression level of truncated retrotransposon copies containing a segment coding for the gag domain could generate an excess of gag protein, which would be required for assembly of virus-like protein and targeting retrotransposable elements to their specific locus. It is interesting to note that in the Nonaut group the truncated copies (Nonaut-1.1, -4.1, and -6.1/-6.2) have a relative transcriptional activity per copy comparable to the complete ones (Nonaut-3 and -5) (Table 1). In addition, four out of the five truncated copies code for a protease domain (Fig. 2C), suggesting that a protease activity may be necessary for correct processing of gag. Interestingly, we also found *S. mansoni* transcripts for a shorter truncated version of Nonaut-5 LTR retrotransposon, which we named Nonaut-5.1 (Figs. 2A and C). The latter lacks the RT coding sequence (2814 bp) that is replaced by a stretch of 552 bp of diverse sequence.

It is interesting to note that two retrotransposons related to Nonaut-5 and -3 were found in *Anopheles gam-*

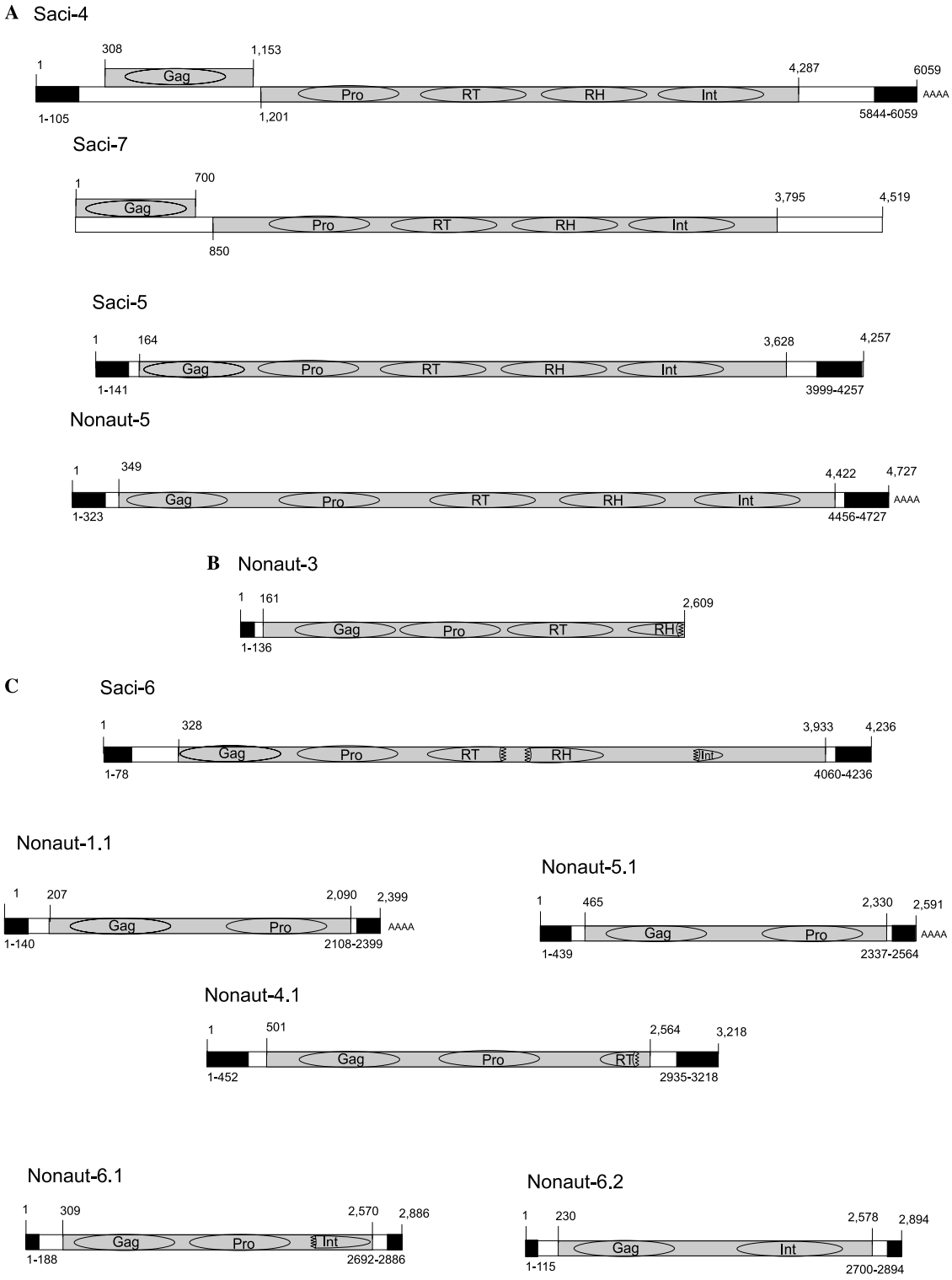

Fig. 2. Diagrammatic representation of new *S. mansoni* LTR retrotransposons. Representation of the sequences obtained by EST reconstruction. (A) Complete retrotransposons with ORFs containing all characteristic domains of LTR retrotransposons. (B) Retrotransposon partially reconstructed with an incomplete ORF that lacks the integrase domain. (C) Truncated retrotransposons displaying LTRs at both extremities but lacking several characteristic domains of LTR retrotransposons. Each rectangle represents one retrotransposon; white portion of the rectangle indicates the non-coding region of retrotransposon. Gray rectangle represents a deduced ORF, and those above the original rectangle indicate an ORF on a different reading frame. Black rectangles represent LTRs and the range indicated below them corresponds to the location of these repeats. Each ellipse represents a different protein domain (Pro, protease; RT, reverse transcriptase; RH, RNase H; and Int, integrase). Full ellipses represent complete domains and broken ellipses indicate partially represented domains. Numbers above the rectangles indicate the position along the sequence in base pairs. Four A's after the rectangle indicate that the reconstruction was anchored at the 3′-end on an EST with a poly(A) tract.

Table 1
Eighteen new retrotransposons of *S. mansoni*

Transposon name	Proposed *Sm* group[a]	Size (bp)	Gene index[f]	Estimated copy number in the genome[g]	Relative transcriptional activity (per copy)[h]
LTR retrotransposons					
Saci-4	Boudicca/Saci-3	6059	0.009	80–800	3.30
Saci-7	Boudicca/Saci-3	4519	0.001	8–80	11.30
Saci-5	Saci-5	4257	0.007	60–600	5.40
Nonaut-3	Fugitive	2607[c]	0.003	29–290	3.00
Nonaut-5	Fugitive	4727[d]	0.013	120–1200	1.38
Nonaut-1.1	Fugitive[b]	2,399	0.005	50–500	11.82
Nonaut-4.1	Fugitive[b]	3218	0.057	490–4900	0.36
Nonaut-6.1	Fugitive[b]	2866	0.012	110–1100	4.80
Saci-6	Saci-1[b]	4236	0.003	20–200	9.87
Non-LTR retrotransposons					
Perere-2	SR1	4544	0.056	480–4800	1.15
Perere-5	SR1	5057	0.042	360–3600	0.63
Perere-6	SR1	4300	0.008	70–700	2.41
Perere-4	Perere	5111	0.012	100–1000	1.33
Perere-7	Perere	4327	0.046	400–4000	0.33
Perere-8	Perere-8	1409[c]	0.018	150–1500	1.43
Perere-3	SR2	3327	0.284	2400–24,000	0.83
Perere-9	Perere-9	4394[e]	—	—	—
Penelope-like element (PLE) retrotransposon					
Perere-10	Cercyon	2258[c,e]	0.022	190–1900	0.21

[a] Groups were defined by phylogenetic analysis of the RT domain of *S. mansoni* retrotransposons. Name of the group is the name of retrotransposons first described within the group.
[b] Since this element lacks the RT domain, proposed group classification was based on sequence similarity and arrangement of other domains present in the element, rather than on direct observation of a phylogeny tree.
[c] Represents the longest reconstructed copy, however containing a partial sequence that codes for the complete RT domain and some additional incomplete domain.
[d] Represents the complete copy of the sequence. Part of this sequence was obtained by RACE, rather than by EST reconstruction.
[e] A genomic clone of this sequence was obtained by PCR in addition to the copy reconstructed from ESTs.
[f] Calculated as described in [5], i.e., the number of hits of the retrotransposon sequence against genomic DNA BAC-end sequences available in GenBank divided by the retrotransposon length.
[g] Estimated by taking the range of the number of copies of Boudicca in the *S. mansoni* genome [7] as a reference and using the gene index factors to calculate the copy number ranges for the other retrotransposons.
[h] Calculated as the number of EST transcripts divided by the transposon length and normalized in relation to SR2 which was taken as 1.0, as described in [5].

biae (ENSANGP18342 protein, Accession No. XP_308479.1) and *Caenorhabditis elegans* (2H728 protein, Accession No. NP_495557.1) (Fig. 1A) for which truncated copies with an intact gag domain were described (*A. gambiae* XP_319741.1; *C. elegans* NP_492520.1 and NP_494296.1) in addition to the full-length copies, suggesting a recurrent pattern in these related retrotransposons.

The non-LTR families

Most of the *S. mansoni* non-LTR-retrotransposons belong either to the RTE or CR1 families, which along with the Jockey family are considered the newest lineages of non-LTR retrotransposons [29]; the abundance of these two families in *S. mansoni* is similar to that described in *C. elegans* [30,31], indicating a possible predisposition of both organisms to facilitate the dispersion of these families.

Non-LTR retrotransposons of SR1 group have a PHD domain

We were able to reconstruct an additional portion of 1669 bp at the 5′ end from the previously described *S. mansoni* non-LTR SR1 retrotransposon [9]. The additional segment encodes an endonuclease and an ORF containing a PHD domain, which is a zinc-finger-like motif found in nuclear proteins thought to be involved in chromatin-mediated transcriptional regulation (data not shown). It has recently been described that CR1 retrotransposons of *A. gambiae* and *Drosophila melanogaster* code for PHD domains, which are proposed to promote efficient transposition and to minimize potentially harmful insertions [32]. In fact, all sequences that clustered as members of the SR1 group of the CR1 family of non-LTR retrotransposons exhibit two ORFs and the first ORF of every retrotransposon codes for a PHD domain (Fig. 3). Interestingly, the other retrotransposons of

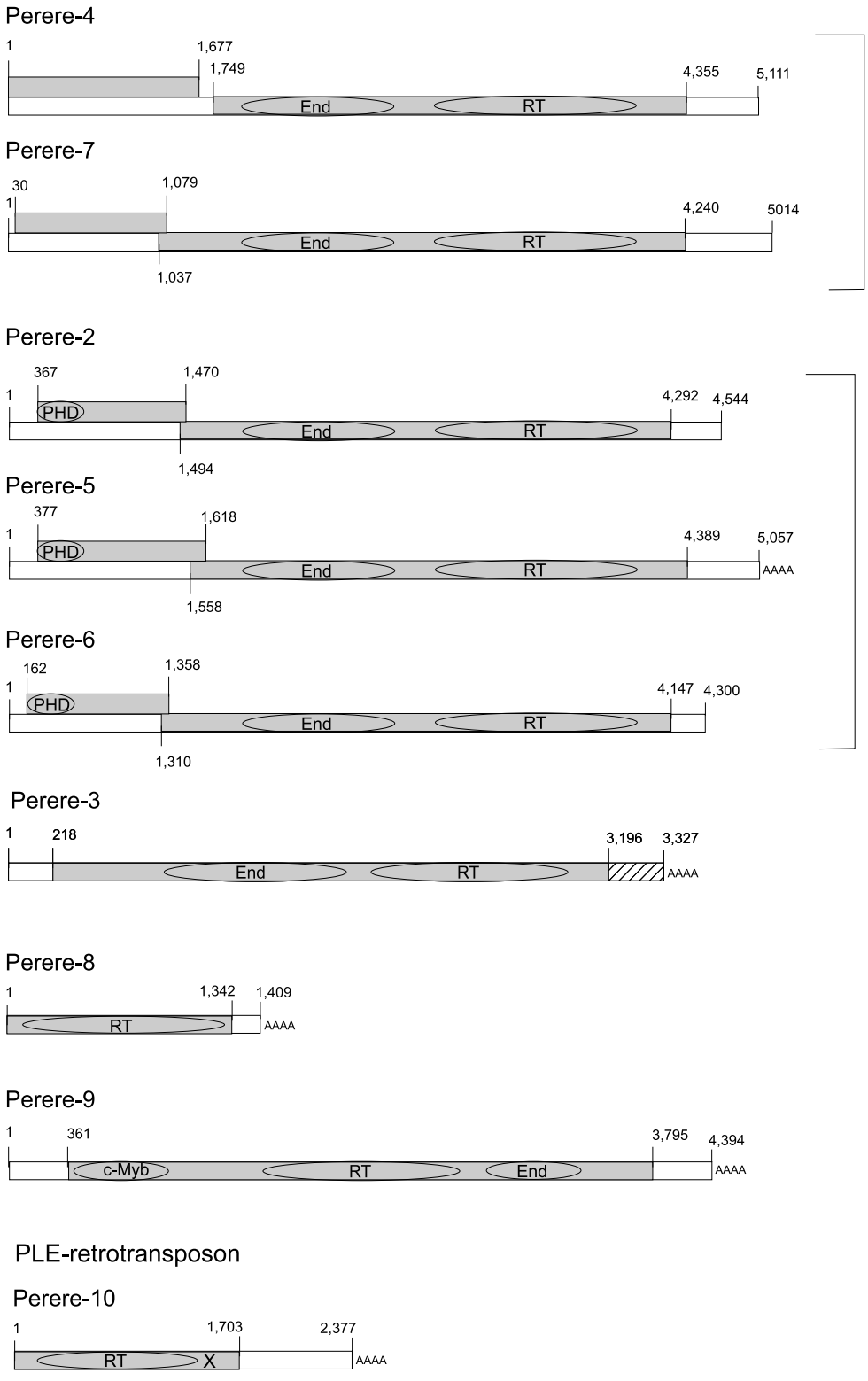

Fig. 3. Diagrammatic representation of new *S. mansoni* non-LTR and PLE retrotransposons. Each rectangle represents one retrotransposon sequence obtained by EST reconstruction. White portion of the rectangle indicates the non-coding region of retrotransposon. Gray rectangle represents a deduced ORF, and those above the original rectangle indicate an ORF on a different reading frame. Each ellipse represents a different protein domain (Pro, protease; RT, reverse transcriptase; RH, RNase H; and Int, integrase). Full ellipses represent complete domains and broken ellipses indicate partially represented domains. Numbers indicate the positions along the sequence in base pairs. Perere-8, -7, and -10 have partial ORFs. Four A's after the rectangle indicate that the reconstruction was anchored at the 3′-end on an EST with a poly(A) tract. The hatched rectangle indicates a variable region. X indicates that stop codons are present but another ORF follows on the same frame.

CR1 family in the Perere group (Perere, Perere-4 and -7) did not exhibit PHD domains in the deduced proteins coded by their ORFs (Fig. 3), providing support for a division of *S. mansoni* retrotransposons into two groups within the CR1 family.

A 3′ terminus was detected in the reconstructed copies of Perere-5 and the previously described retrotransposon Perere that contains a repeat motif $(ATT)_X$. This was confirmed by alignment with several BAC end sequences that contain the extremity of these retrotransposons with the same repeated motif (data not shown). For Perere-5, X varied from two to fifteen repeats and occurred just after a polyadenylation signal AATAAA. In addition, the mRNA message of Perere-5 displayed a poly(A) tract just after the repeats, which was not present in BAC end sequences. Although the presence of such repeats is common to non-LTR retrotransposons of the CR1 family, they were not described in SR1 retrotransposons, which instead have an $(ACC\ ATGG)_2$ motif at their 3′ extremity [9].

Perere-9 is a novel R2 non-LTR retrotransposon

We were able to obtain a genomic clone of Perere-9 from a PCR using primers designed from EST sequences. This clone exhibits a mutation that inserts a stop codon 1623 bases downstream from its start and codes for a smaller ORF when compared to the reconstructed sequence.

As noted above, Perere-9 belongs to the R2 family of non-LTR retrotransposons based on the phylogeny analysis of the deduced RT domains (Fig. 1B). This is the first R2 family retrotransposon described in *Schistosoma*. Alignment of the R2 protein coded by the reconstructed copy of Perere-9 along with several R2 proteins from retrotransposons of other species (Supplementary Fig. 2) showed that it exhibits the conserved amino-terminal CCHH and c-myb domains [33], and a characteristic carboxyl-terminal endonuclease domain with a CCHC/PD..D motif [29] previously described for the R2, R4, and CRE families. These three oldest lineages of non-LTR retrotransposons usually encode such endonucleases with an active site similar to that of certain restriction enzymes and show target site specificity for unique locations in the genome [34]. Based on this observation we hypothesize that Perere-9 may also insert into specific DNA sites in the genome. In fact, we found an *S. mansoni* EST (Accession No. CF497869) that has both sequences of Perere-9 and of ribosomal intergenic spacer DNA (Accession No. AJ223842), which probably originated from an immature rRNA message with an insertion of Perere-9. Further confirmation is obtained by a contig of whole genome shotgun (WGS) sequences (2303595.c002142339.Contig1) from the Sanger Institute preliminary genome assembly that displays sequences both of Perere-9 and of the ribosomal intergenic spacer DNA adjacent to each other. If this were true, the specificity of insertion into ribosomal intergenic spacer DNA would be analogous to that described for the Dong retrotransposon of *Bombyx mori* [29] of the R4 family. It is interesting to note that a blastn query of Perere-9 sequence against the database of 27,064 *S. mansoni* genomic DNA BAC-end sequences resulted in no hit, suggesting that this transposon might be present at a very low copy number in the parasite's genome.

Perere-3 transcripts exhibit a variable 3′-end

We failed to reconstruct Perere-3 very 3′-end, because multiple possibilities for reconstruction were detected, without a predominant message in this region. We decided to use an alternative clustering approach, as described in the Supplementary methods, in order to verify if a prevalent sequence could be detected. In brief, a Blastn search against the *S. mansoni* EST database was performed using a 270 bp fragment at the most 3′-end portion of conserved sequence of Perere-3 as query (bases 2927 to 3196 of reconstructed sequence). This search retrieved 166 ESTs with an alignment score above 100, each having an adjacent 3′-end downstream sequence of at least 50 bp. After masking the Perere-3 conserved sequence of all retrieved ESTs, the remaining unmasked 3′-end sequences were clustered using CAP3. The output was further grouped by running a Blastn of each resulting contig and singlet against each other. This methodology produced 29 different groups and 81 singlets, the largest group being composed of only 8 sequences, thus representing only 5% of the ESTs. The same approach was used with *S. mansoni* genomic DNA BAC-end sequences as database and retrieved 322 hits; again, grouping produced 38 different groups and 213 singlets, the largest group being composed of 13 genomic DNA sequences.

As a control for our CAP3/BLASTN clustering approach, we used another region of Perere-3 totally contained within the conserved region both as query (bases 2267–2536) and as the unmasked region to be clustered (bases 2537–3196). A search against the EST database retrieved 91 ESTs and grouping generated only one group composed of 84 ESTs, thus representing 92% of the retrieved sequences, its consensus sequence being similar to the sequence reconstructed by us for this conserved segment.

Overall, the above analysis points out that no predominant sequence exists at the 3′-end of Perere-3 downstream of base 3196, either in the genomic or in the transcribed copies. In order to further document that diverse copies were transcribed, we extracted $poly(A)^+$ mRNA from adult worms and performed a 3′-end RACE experiment for Perere-3, followed by sequencing of six different clones. Multiple alignment

Fig. 4. Sequence alignment of six different 3′ end clones for Perere-3 obtained by RACE. Alignment was performed with Clustal X program. Columns presenting an asterisk are those in which a nucleotide is conserved in all six clone sequences. Shaded nucleotides are those of the predicted early stop codon for four of the sequences. Box represents the conserved putative stop codon for the longest ORF. Numbers indicate the position relative to the full-length reconstructed sequence.

(Fig. 4) showed that these six clones had a highly conserved region extending up to base 3196, immediately followed by divergent 3′-ends of different lengths plus a poly(A) tail. These copies have point mutations and deletions in their putative ORFs that caused frame shifts and originated early stops in some of the deduced translated sequences. Nevertheless, a conserved potential stop codon (TAA) can be identified at bases 3194–3196 in all copies (Fig. 4, boxed). Divergence occurs at the same point as that verified in individual ESTs previously sequenced in the high-throughput transcriptome sequencing project [25] and in BAC-end genomic sequences from GenBank (not shown).

One possible explanation is that the ESTs and RACE sequences originated from messages transcribed by active copies of Perere-3 that are co-expressed with a variable 3′-end near the poly(A) tail. In that scenario, Perere-3 would play an important role in *S. mansoni* genome dynamics, since it probably can transfer different DNA sequences to new genomic sites, similar to those proposed for retrotransposon L1 in humans [35]. RTE elements have been previously described to have a variable 3′-end [30], however these variable ends were few bases long and presented low complexity. In contrast, the 3′-end sequences described here are from 40 up to 450 base pairs long and apparently do not contain repetitive motifs (Fig. 4). We have no evidence for any possible mechanism that would be involved in the process and therefore cannot provide further confirmation to this hypothesis.

An alternative explanation is that the full-length retrotransposon extends only up to the conserved 3′-end portion, and the variable region that we have detected as transcribed, in fact, arises from read-through activity of neighbor genes. If this were the case, Perere-3 would exhibit very remarkable unique characteristics in relation to other retrotransposons: (1) its 3′-end would be located at the same point of its probable ORF stop codon; and (2) most of its copies that reside near bona fide transcribed genes would be non-truncated, since these supposed read-through activities predominantly revealed intact non-truncated copies. This is a very unlikely event, since a large number of truncated copies of retrotransposons are known to be present in the genome of eukaryotes and would show up in transcripts originated from read-through activities. Rather, the predominance of non-truncated copies in the EST database suggests that the former explanation is the most likely one.

The relative transcriptional activity of the retrotransposons

Relative transcriptional activity of the retrotransposons was estimated (Table 1), using the approach described previously [5]; the relative transcriptional activity is calculated as the number of EST transcripts divided by the transposon length and normalized in relation to SR2 [5]. For each described *S. mansoni* retrotransposon, plotting the relative transcriptional activity as a function of the gene index, which is a parameter related to the transposon copy number [5] (i.e., the number of hits to genomic DNA BAC-ends divided by the transposon length) allowed us to discern two distinct populations (Fig. 5). The first has a low genomic copy number but high transcriptional activity; in contrast, the second displays high genomic copy number and low transcriptional activity. Such data provide additional support for the hypothesis that two populations

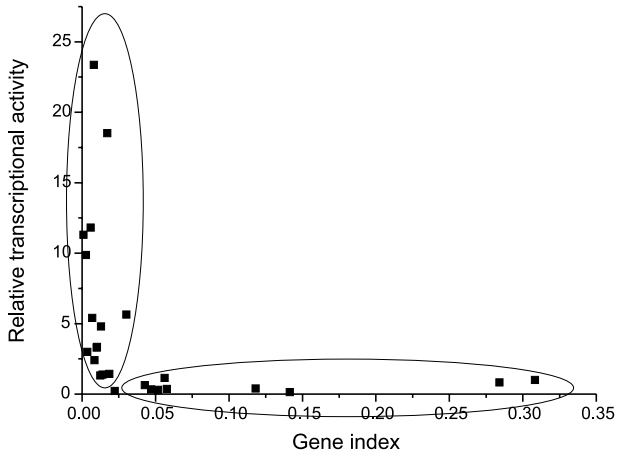

Fig. 5. Comparison of gene index and relative transcriptional activity. Data regarding relative transcriptional activity as a function of gene index for 24 different *S. mansoni* retrotransposons were plotted. Data were extracted from Table 1 for the elements described here and from DeMarco et al. [5] for the previously described elements. Ellipses highlight two populations with marked different characteristics.

with very distinct traits coexist in the *S. mansoni* genome [5]. In addition, this trait is the opposite to what would be expected if retrotransposon transcripts originated from read-through activities of neighbor genes. In this case, transcription of retrotransposon elements would be a random event, being proportional to the copy number of each element; a similar transcriptional activity per copy would be expected for all retrotransposons. In contrast, we observed a very distinct activity for each retrotransposon (Fig. 5), suggesting that distinct signaling is driving the transcription of some of them.

The individual retrotransposon ESTs obtained in the high-throughput *S. mansoni* transcriptome sequencing project [25] that are analyzed here do have on average a high identity (91–98%) with respect to the corresponding full-length reconstructed sequence for each of the 18 new retrotransposons (Supplementary Table 1). Most important, coverage of individual EST sequences by the corresponding full-length reconstructed sequence is high on average (74–94%), especially considering that single-pass sequences are prone to present base calling errors and regions of low quality sequencing. Again, high coverage suggests that the messages from which each of the ESTs originate do appear to contain only retrotransposon sequence derived from a true transposon transcription rather than containing a chimeric read-through transcript with retrotransposon plus neighbor gene sequence.

Conclusions

The present study allowed us to access the message of transcribed copies of several new retrotransposon elements so far undetected in the *S. mansoni* EST database as well as to reconstruct prototypes of their full-length copies. Conservation of their traits and evidence that they are transcribed suggest that these transposons are the most likely elements to play a role in the shaping of *S. mansoni* genome. Identification of such full-length uncorrupted messages as described here is the first step to characterize their active copies; further tests of the putative encoded proteins through expression in heterologous systems and in vivo experiments of transposition are need in order to confirm their postulated activity. Further characterization of such elements is warranted since they may open the possibility to use these elements as tools to manipulate *Schistosoma* genome biology.

Finally, the *S. mansoni* genome is currently being sequenced [36] as a collaborative project between TIGR (http://www.tigr.org/tdb/e2k1/sma1/) and the Sanger Institute (http://www.sanger.ac.uk/Projects/S_mansoni/). The genome is estimated to be large in size (270 Mbp) and with a considerable amount (~40%) of repetitive sequence [2]. The most recently released assembly shows that the genome sequence is still fragmented into 70,714 contigs. Knowing the full-length sequence of 18 new retrotransposons, as obtained by an independent method such as the reconstruction from transcribed sequences described in the present work, could help in the assembly of the parasite's genome sequence.

Acknowledgments

This work was financed by Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) and by the Brazilian Ministry of Science and Technology, Conselho Nacional de Desenvolvimento Científico e Tecnológico (MCT, CNPq). We thank Thiago M. Venâncio for his help in parsing BLAST results.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2005.05.080.

References

[1] WHO, TDR Strategic Direction for Research: Schistosomiasis, Ed., World Health Organization, Geneve. 2002.
[2] A.J. Simpson, A. Sher, T.F. McCutchan, The genome of *Schistosoma mansoni*: isolation of DNA, its size, bases and repetitive sequences, Mol. Biochem. Parasitol. 6 (1982) 125–137.
[3] T. Laha, P.J. Brindley, C.K. Verity, D.P. McManus, A. Loukas, Pido, a non-long terminal repeat retrotransposon of the chicken repeat 1 family from the genome of the oriental blood fluke *Schistosoma japonicum*, Gene 284 (2002) 149–159.

[4] R. Hull, Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses, Arch. Virol. 146 (2001) 2255–2261.
[5] R. DeMarco, A.T. Kowaltowski, A.A. Machado, M.B. Soares, C. Gargioni, T. Kawano, V. Rodrigues, A.M.B.N. Madeira, R.A. Wilson, C.F.M. Menck, J.C. Setubal, E. Dias-Neto, L.C.C. Leite, S. Verjovski-Almeida, Saci-1, -2and -3 and Perere, four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*, J. Virol. 78 (2004) 2967–2978.
[6] T. Laha, A. Loukas, D.J. Smyth, C.S. Copeland, P.J. Brindley, The fugitive LTR retrotransposon from the genome of the human blood fluke, *Schistosoma mansoni*, Int. J. Parasitol. 34 (2004) 1365–1375.
[7] C.S. Copeland, P.J. Brindley, O. Heyers, S.F. Michael, D.A. Johnston, D.L. Williams, A.C. Ivens, B.H. Kalinna, Boudicca,a retrovirus-like long terminal repeat retrotransposon from the genome of the human blood fluke *Schistosoma mansoni*, J. Virol. 77 (2003) 6153–6166.
[8] C.S. Copeland, V.H. Mann, M.E. Morales, B.H. Kalinna, P.J. Brindley, The Sinbad retrotransposon from the genome of the human blood fluke, *Schistosoma mansoni*, and the distribution of related Pao-like elements, BMC Evol. Biol. 5 (2005) 20.
[9] A.C. Drew, P.J. Brindley, A retrotransposon of the non-long terminal repeat class from the human blood fluke *Schistosoma mansoni*. Similarities to the chicken-repeat-1-like elements of vertebrates, Mol. Biol. Evol. 14 (1997) 602–610.
[10] A.C. Drew, D.J. Minchella, L.T. King, D. Rollinson, P.J. Brindley, SR2 elements, non-long terminal repeat retrotransposons of the RTE-1 lineage from the human blood fluke *Schistosoma mansoni*, Mol. Biol. Evol. 16 (1999) 1256–1269.
[11] I.R. Arkhipova, K.I. Pyatkov, M. Meselson, M.B. Evgen'ev, Retroelements containing introns in diverse invertebrate taxa, Nat. Genet. 33 (2003) 123–124.
[12] C. Feschotte, Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences, Mol. Biol. Evol. 21 (2004) 1769–1780.
[13] M.G. Ivanchenko, J.P. Lerner, R.S. McCormick, A. Toumadje, B. Allen, K. Fischer, O. Hedstrom, A. Helmrich, D.W. Barnes, C.J. Bayne, Continuous in vitro propagation and differentiation of cultures of the intramolluscan stages of the human parasite *Schistosoma mansoni*, Proc. Natl. Acad. Sci. USA 96 (1999) 4965–4970.
[14] B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of repetitive DNA in eukaryotes, Nature 371 (1994) 215–220.
[15] J.A. Yoder, C.P. Walsh, T.H. Bestor, Cytosine methylation and the ecology of intragenomic parasites, Trends Genet. 13 (1997) 335–340.
[16] M.T. Clegg, M.L. Durbin, Flower color variation: A model for the experimental study of evolution, Proc. Natl. Acad. Sci. USA 97 (2000) 7016–7023.
[17] A.D. Long, R.F. Lyman, A.H. Morgan, C.H. Langley, T.F. Mackay, Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete–scute complex are associated with variation in bristle number in *Drosophila melanogaster*, Genetics 154 (2000) 1255–1269.
[18] E.W. Ganko, V. Bhattacharjee, P. Schliekelman, J.F. McDonald, Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution, Mol. Biol. Evol. 20 (2003) 1925–1931.
[19] V.V. Kapitonov, J. Jurka, The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor, J. Mol. Evol. 48 (1999) 248–251.
[20] D.T. Scholes, A.E. Kenny, E.R. Gamache, Z. Mou, M.J. Curcio, Activation of a LTR-retrotransposon by telomere erosion, Proc. Natl. Acad. Sci. USA 100 (2003) 15736–15741.
[21] T.A. Morrish, N. Gilbert, J.S. Myers, B.J. Vincent, T.D. Stamato, G.E. Taccioli, M.A. Batzer, J.V. Moran, DNA repair mediated by endonuclease-independent LINE-1 retrotransposition, Nat. Genet. 31 (2002) 159–165.
[22] E. Casacuberta, M.L. Pardue, HeT-A elements in *Drosophila virilis*: retrotransposon telomeres are conserved across the *Drosophila genus*, Proc. Natl. Acad. Sci. USA 100 (2003) 14091–14096.
[23] V.V. Kapitonov, J. Jurka, Rolling-circle transposons in eukaryotes, Proc. Natl. Acad. Sci. USA 98 (2001) 8714–8719.
[24] Z. Ivics, P.B. Hackett, R.H. Plasterk, Z. Izsvak, Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells, Cell 91 (1997) 501–510.
[25] S. Verjovski-Almeida, R. DeMarco, E.A. Martins, P.E. Guimaraes, E.P. Ojopi, A.C. Paquola, J.P. Piazza, M.Y. Nishiyama, J.P. Kitajima, R.E. Adamson, P.D. Ashton, M.F. Bonaldo, P.S. Coulson, G.P. Dillon, L.P. Farias, S.P. Gregorio, P.L. Ho, R.A. Leite, L.C. Malaquias, R.C. Marques, P.A. Miyasato, A.L. Nascimento, F.P. Ohlweiler, E.M. Reis, M.A. Ribeiro, R.G. Sa, G.C. Stukart, M.B. Soares, C. Gargioni, T. Kawano, V. Rodrigues, A.M. Madeira, R.A. Wilson, C.F. Menck, J.C. Setubal, L.C. Leite, E. Dias-Neto, Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*, Nat. Genet. 35 (2003) 148–157.
[26] T. Jacks, M.D. Power, F.R. Masiarz, P.A. Luciw, P.J. Barr, H.E. Varmus, Characterization of ribosomal frameshifting in HIV-1 gag-pol expression, Nature 331 (1988) 280–283.
[27] V. Karacostas, E.J. Wolffe, K. Nagashima, M.A. Gonda, B. Moss, Overexpression of the HIV-1 gag-pol polyprotein results in intracellular activation of HIV-1 protease and inhibition of assembly and budding of virus-like particles, Virology 193 (1993) 661–671.
[28] A. Atwood, J.H. Lin, H.L. Levin, The retrotransposon Tf1 assembles virus-like particles that contain excess Gag relative to integrase because of a regulated degradation process, Mol. Cell. Biol. 16 (1996) 338–346.
[29] J. Yang, H.S. Malik, T.H. Eickbush, Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements, Proc. Natl. Acad. Sci. USA 96 (1999) 7847–7852.
[30] H.S. Malik, T.H. Eickbush, The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs, Mol. Biol. Evol. 15 (1998) 1123–1134.
[31] I. Marin, P. Plata-Rengifo, M. Labrador, A. Fontdevila, Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode *Caenorhabditis elegans*, Mol. Biol. Evol. 15 (1998) 1390–1402.
[32] V.V. Kapitonov, J. Jurka, The esterase and PHD domains in CR1-like non-LTR retrotransposons, Mol. Biol. Evol. 20 (2003) 38–46.
[33] W.D. Burke, H.S. Malik, J.P. Jones, T.H. Eickbush, The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods, Mol. Biol. Evol. 16 (1999) 502–511.
[34] H.S. Malik, T.H. Eickbush, NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*, Genetics 154 (2000) 193–203.
[35] J.V. Moran, R.J. DeBerardinis, H.H. Kazazian Jr., Exon shuffling by L1 retrotransposition, Science 283 (1999) 1530–1534.
[36] N.M. El-Sayed, D. Bartholomeu, A. Ivens, D.A. Johnston, P.T. LoVerde, Advances in schistosome genomics, Trends Parasitol. 20 (2004) 154–157.