SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs

Serap Aksoy[1,2*], Suzanne Williams[1], Sandy Chang[+] and Frank F.Richards[1,2]
[1]Yale MacArthur Center for Molecular Parasitology and [2]Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06510, USA

Received November 17, 1989; Revised and Accepted January 11, 1990 EMBL accession no. X17078

ABSTRACT

We have characterized a retrotransposon in *Trypanosoma brucei gambiense* uniquely associated with the spliced-leader (SL) RNA gene cluster (Spliced Leader Associated Conserved Sequence, SLACS). There are nine copies of SLACS and DNA sequence analysis of one shows the hallmarks of Line-1 like elements. SLACS has generated a 49 bp target DNA duplication at its insertion site and its 3′-end is preceeded by a poly(A) stretch. Two putative open reading frames (ORFs) span 75% of the element. ORF1 has CysHis motif associated with the retroviral gag polypeptide while ORF2 shows homology with reverse transcriptase sequences. Its 5′-end contains a repeated segment of a 185 bp that varies in copy number in different SLACS insertions. Retrotransposon-like sequences inserted into the SL-RNA genes occur in several hemoflagellates. These elements may represent a related family which has maintained its target site specificity.

INTRODUCTION

The presence of insertion elements showing homology with mammalian long interspersed nuclear elements (LINEs) (reviewed 1,2) has been reported in many genomes (3–12). These sequences lack the long terminal repeats (LTRs) of retrovirus-like elements. However, they have generated target DNA duplications at their insertion sites. There is a characteristic A-rich region or poly(A) tail at the 3′-ends of such elements. Another feature shared by these sequences is the presence of two open-reading frames (ORFs) that span most of the coding regions. The amino acid sequence analysis of the longer ORF (ORF2) shows significant homology with reverse transcriptases of retroviruses (13,14); while the shorter ORF (ORF1) often bears homology with the nucleic acid binding domains of gag gene polypeptides (15). There are also LINE-1 like elements which contain only one long ORF showing homology with reverse-transcriptase-like sequences (4,11,12). While the majority of these elements appear to be randomly distributed throughout the genome, examples of target site specific insertion events have also been reported as in the case of the R2Bm element in *Bombyx mori* (16). This suggests that endonuclease sequences recognizing these insertion site regions might also be coded for by these ORF(s).

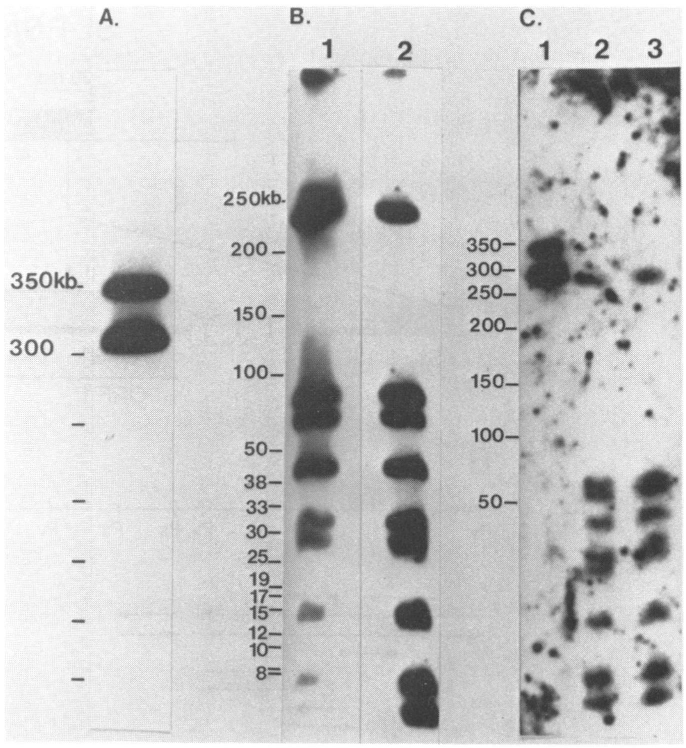

Figure 1: Genomic Organization of SLACS and SL-RNA genes. A. Pulse-field gel electrophoresis (PFG) analysis of EcoRI cleaved *T.b.gambiense* chromosomal DNA hybridized with SL-RNA probe (described in Materials and Methods). For separation of fragments, a pulse time of 40 seconds and 170V was used for 36 hours. Ligated lambda DNA is used as size marker. B. *T.b.gambiense* chromosomal DNA cleaved with the restriction enzyme BglII was separated by PFG-Hexagonal Array using 10 second pulse time at 170 V for 30 hours.In Lane 1, the hybridization probe was SL-RNA gene. In Lane 2 the same Southern blot was hybridized to SLACS probe, pSB1. C. PFG analysis of EcoRI(Lane 1), BamHI(Lane 2) and SalI(Lane 3) cleaved *T.b.gambiense* chromosomal DNA separated using a pulse time of 15 seconds for the initial 15 hours followed by 5 seconds for another 15 hours at 170 V. The hybridization probe was SLACS specific pSB1.

* To whom correspondence should be addressed

[+] Present address: Rockefeller University, 1230 York Ave, New York, NY 10021, USA

Previously we reported that the spliced-leader (SL) RNA genes of the African trypanosome, *T.b.gambiense*, contain interrupting sequences (17). Similar SL-RNA gene interrupting sequences in *T.b.gambiense* were also independently reported (18). There are approximately 300 copies of SL-RNA genes organized in tandem units into one or two clusters. Our results indicated that 9 of these SL-RNA genes are interrupted by insertion elements between the 11th and 12th nucleotide from the 5'-end of their coding sequences. Based on a fine mapping Southern blot analysis, we concluded that the overall organization of each insertion element is the same. Thus, we refer to this element as Spliced-Leader Associated Conserved Sequence (SLACS). Because the element is flanked by a duplicated 49 bp target DNA sequence at its site of insertion and because there is a long poly(A) tail associated with its 3'-end, we concluded that it was a retrotransposon.

In this paper, we report the complete nucleotide sequence of one of these SLACS which suggests that it belongs to the family of non-LTR containing elements exhibiting pol and gag polypeptide homologies. More recently, another insertion sequence within the SL-RNA genes has been observed in *Crithidia fasciculata*, a monogenetic mosquito trypanosomatid at some evolutionary distance from *T.b.gambiense* (19). The overall organization of this element also resembles the non-LTR containing retrotransposons. The *C.fasciculata* element has also been inserted into the analogous site in the same target DNA sequence as SLACS. We discuss the evolutionary implications of these findings.

MATERIALS AND METHODS

Trypanosome Stocks and Genomic Library Screening

T.b.gambiense cloned variant antigen types of the Texas trypanozoon antigen type (Txtat) serodeme were used (20). This variant was a gift of Dr.John R.Seed, University of North Carolina. The Txtat I DNA library constructed in the EMBL3 bacteriophage vector was a gift from Dr. Christian Tschudi (20). Screening of genomic clones containing the SL-RNA genes and retrotransposon sequences, phage DNA isolation and Southern blot analysis have been described previously (17). The preparation of trypanosome blocks for pulse field gel (PFG) electrophoresis analysis has also been described (17).

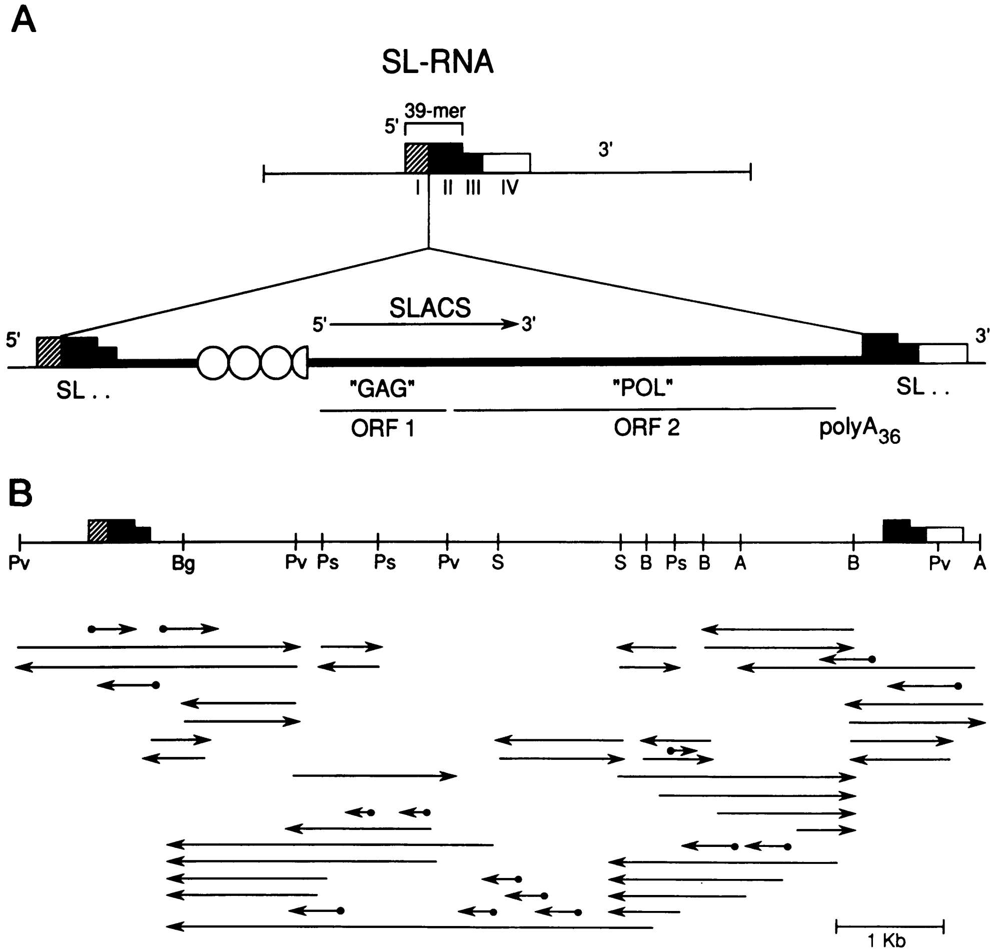

Figure 2: Organization of SLACS inserted into one SL-RNA gene A. Schematic diagram showing the location of SLACS within a 1.4 kb SL-RNA gene unit in *T.b.gambiense*. The orientation of transcription for SL-RNA and SLACS sequences is shown by the 5'to 3'arrow. SL-RNA coding sequence is divided into 4 regions. The 5'end 39-mer SL sequence is composed of boxes I and II. Box I contains the first 11 nucleotides, Box II has the rest of the 27 bases. Box III extends from nucleotides 39 to 60 and Box IV from 60 to the end of the gene. SLACS sequence is denoted by the thicker line and interrupts the SL unit at position 11. Boxes II and III are found at both ends of SLACS and represent the duplicated target sequence. Preceeding the site of insertion at the 3'-end,the stretch of poly(A) is shown. The 3 full and one half circle in SLACS correspond to the 185 bp repeated sequence that varies in copy number in different SLACS. The lines labelled ORF1 and ORF2 show the position of the two open reading frames. B. Partial restriction map of SLACS inserted into one SL-RNA gene present in the phage clone Q107. Abbreviations for the restriction sites are as follows: B,BamHI; S,SalI; P,PstI; Pv,PvuII; A,ApaI; Bg,BglII. Shown below the restriction map are the direction and extent of the sequence determinations. The complete sequence of both strands was determined.

```
                                        1
1     aactaacgctattattagaacagtttctgtac? atattggtatgagaagctcccagtaggAATTATCCGTACTTGGGGTCAATATTCGGGAAGAGAAAGA   100
101   AGTAAGAAATCGCTGCGTTTTATGATATCGATAGGAAAGGAA?GAAAATCCTCAAAACCACAAAAAGTCTTGTTTTGGGGGTTCGAACCCGGACCTCCAA   200
201   AACACAAACACAATAGGAGAGGAGTGTTGCCAGTTGGGCTAT?TCGACAAATCGGAGAAAAAGGAGAAAATTTATGCTTAGATGAAATATACCAAATCGT   300
301   ACACATCGAAGAGAAAAAACATAGGTGTACAAACGGTGCACACGTAGAAAAAAGTAGCGGAATTGATAGATCTCAGGTAAATAAACCCATTACAGAAATC   400
401   AGCGGATCAGTTTCTTTCTTGGGACAGAATATCGGGAGGTTTGATGTCAAAAAGTCACCCGAAAATGAATATTTTGAAAAAATCAACTGTGGTCAAATCG   500
                                                    2
501   ACATAACGACCACAAAAAATTGTCGGCTACGCAGACTTTCGTAGCG GATTTGAAAGGGATCGACCAGCCCAATGCATGAGTAACAACGGTTACTGCGACA   600
601   AAGAGGTCACAAAAAAATTAGGTGCTGCCGAACCTAAAATTTTTTCAAAAAATGGAAATTTTGAAAAAATCAACTGGTCAATCGACATAACGACCACAAA   700
701   AAATTGTCGGCTAGCGGAAGGCTTTCGTAGCGCATTTGAAAGCTCGACCAGCCCAATGCATGAGTAACAACGGTTACTGCGACAAAGAGGTCACAAAAAA   800
801   ATTAGGTGCTGCCGAACCTAAAATTTTTTCAAAAAATGGAAATTTTGAAAAAATTAACCGTGGTCAAATCGACATGACGACCACAAAAAATTGTCGTACG   900
901   GCAGACTTTCGTAGCGCATTTGAAAGGCTCGACCAGCCCAATGCATGAGTAACAACGGTTACTGCGACAAAGAGGTCACAAAAAAATTAGGTGCTGCCGA   1000
1001  ACCTAAAATTTTTTCAAAAAATGGAAATTTTGAAAAAATTAACCGTGGTCAAATCGACATGACGACCACAAAAAAATTGTCGTACGCAAGACCTTTCGTA   1100
1101  GCGCATTTGAAAGCTCGACCAGCCCAATGCATGAGTAACAACGGTTACTCCGACAAAGAGGTCATAAGAAAATAGGGCGTTGGAAAAGAAAAATGCTCAA   1200
1201  AAAATAAGGAAAAAGAGCAAATTTCGCAGGCCGAAAAAAACGACACAAATGGTTTCTACACAGACTTCCGTAGCGCATTTACACGGAATGGACGAGTAGA   1300
1301  ATTAGGCAATAACAACGGTTATTGTCATAACGAAGCGACAAGAAATATAGAGAGGGGGAAAAAGCCGAAAACATCAAAAATTGATGAAAAAAGGCAAATT   1400
1401  TCCAACATTGCAGCTGGTCGAAAAAGTGATCACTACGAAGATCCGAGTAGTGCATTTTCACGAATTTCAATAGACCATTGCAATGCTCAAATAAAAATTG   1500
1501  GTCGCAAAAAAATGGTTAGAAACTTGAGGTCATCTGAACCTCAAAAAATTTCAAGGTTCCCAAGGCACCATGGGGAAGGAAAAAAGGCGTCTTCTCCGGC   1600
                  M  V  R  N  L  R  S  S  E  P  Q  K  I  S  R  F  P  R  H  H  G  E  G  K  K  A  S  S  P  A
1601  TTTGAACCTGCAGTGGCCAGGACAGAAGCAGGTAGCGTCGGCAGTTGCGCCCGAAATGAGAAAAAGGGCGCCCCCGAAAAACAACAACACAAGCATCCAC   1700
       L  N  L  Q  W  P  G  Q  K  Q  V  A  S  A  V  A  P  E  M  R  K  R  A  P  P  K  N  N  N  T  S  I  H
1701  CGGAACTGTAGGAAACACCTTAGGAAAGAAGGTGATTGGTGGATAGTAGAGGGCGGGCAAAACCACAAGAAAGCCGCCCATTCCCCTCTCAAAACGAAAC   1800
      R  N  C  R  K  H  L  R  K  E  G  D  W  W  I  V  E  G  G  Q  N  H  K  K  A  A  H  S  P  L  K  T  K  P
1801  CGGTAAATAAGGAGGGGAACCGGAAGAAGTACGGCAGACCACCGCGTGAGGTAGAAGGGAAATGGCTTACCTCCCTCATTGCGGCAACGACAGAGGCCGT   1900
        V  N  K  E  G  N  R  K  K  Y  G  R  P  P  R  E  V  E  G  K  W  L  T  S  L  I  A  A  T  T  E  A  V
1901  ATTACGCCAACTGGGGAAGGGAAAATCCCCAACAGCCCACACGAAGTCGAACCAGAGTAGAGTCCCGCTGCAGCACTCTGAGAAAACGGCTAAGGGCATC   2000
       L  R  Q  L  G  K  G  K  S  P  T  A  H  T  K  S  N  Q  S  R  V  P  L  Q  H  S  E  K  T  A  K  G  I
2001  AAATACGCGACGCCCCAACCAAAAAAGAAAGCACATGAGCAGCGGAAAGAGCTCCCCCATTGGCCCCCAACCAAGAGGAGCCACGGCGCTAACAAAGGGC   2100
      K  Y  A  T  P  Q  P  K  K  K  A  H  E  Q  R  K  E  L  P  H  W  P  P  T  K  R  S  H  G  A  N  K  G  Q
2101  AGGGAGCACCAGTAAGAGCCCCTGCGAAGACACAAGGGAAGGGGGAAGAACAACCCCACACAATGCGAACATGGGCACAGGTGGCAGCACCGAAGAAACA   2200
        G  A  P  V  R  A  P  A  K  T  Q  G  K  G  E  E  Q  P  H  T  M  R  T  W  A  Q  V  A  A  P  K  K  Q
2201  AAAGGTGACAAGTAAACCACCCATGGCTCAAAAAAAAAAGGCACAGGGGGAAAAGAAAGGGGCAGCGGCCAACCCCTTCCACTGGGAGCTACAGGTGGAG   2300
       K  V  T  S  K  P  P  M  A  Q  K  K  K  A  Q  G  E  K  K  G  A  A  A  N  P  F  H  W  E  L  Q  V  E
2301  CAGCTTCTCAAGGATGCGGAACAGATACGCGAAAGCATGTACATTCGCTTTCTCCACGTTCGGCAGAGTTGGTGGAGCACATGCTTCAAAGCCCACATGG   2400
      Q  L  L  K  D  A  E  Q  I  R  E  S  M  Y  I  R  F  L  H  V  R  Q  S  W  W  S  T  C  F  K  A  H  M  E
2401  AGTTCCACTGCCCAGTGTGTGGTTTCGCGCACCCGGAAGAAACCATAACAGTAACACACTGCAGGCAGCAACACCCAGGAGGGCCGCCTGATTCCCTACA   2500
                                                 3
        F  H  C  P  V  C  G  F  A  H  P  E  E  T  I  T  V  T  H  C  R  Q  Q  H  P  G  G  P  P  D  S  L  H
2501  CCCTGACAACAACAGGGAATCAGGTGCAGTGTCGCAGGTCCCTCCTGCTCACTTTCGGCGGTGGTGGCTATCCTCTCACATATGGAGGAAGAGAAAAATT   2600
        P  D  N  N  R  E  S  G  A  V  S  Q  V  P  P  A  H  F  R  R  W  W  L  S  S  H  I  W  R  K  R  K  I
2601  TCCACCCTATTATCGCCGAAGCTACCAGTCCCCGAGCAAGGAGACTACCCAAACGCTCGTTCAGGCATTGGGACTGGAGCTCCCCACAGCCCCCATCACT   2700
      S  T  L  L  S  P  K  L  P  V  P  E  Q  G  D  Y  P  N  A  R  S  G  I  G  T  G  A  P  H  S  P  H  H  C
2701  GCGATAGAAGCGCTGGCCCACCATGACCAGATGATCCGGCAGCTGTTGGAGGCACCGCAACGACGGCACGACAATGCGGGCATTGTGGGGTAAATGAGCT   2800
        D  R  S  A  G  P  P  *
2801  AATGGAAGGGCCTTCCAACCCCGGGGAAAATTTGGGGATAATCTCCGTAAACCCCAACACATGCGAGTCGATCGAGCTGACGAATCAGATCCTCGCGAAG   2900
        M  E  G  P  S  N  P  G  E  N  L  G  I  I  S  V  N  P  N  T  C  E  S  I  E  L  T  N  Q  I  L  A  K
2901  ACATATACGACCTCCTGGACAGATATGGAGGTCTATGCACGGGGGAACGGAAGAATATGCACCGCCTACCCCGACGTTGATCACCCACTCCAAGACGTTC   3000
      T  Y  T  T  S  W  T  D  M  E  V  Y  A  R  G  N  G  R  I  C  T  A  Y  P  D  V  D  H  P  L  Q  D  V  R
3001  GGGGAAATGTGTCATCGATATTGGAACGTGGGGAACGAAGCGGTGGGTTTGAAACGCATCCTCCACGAGATCCTGTTACCCCACGGGAACGGCACGGCAA   3100
        G  N  V  S  S  I  L  E  R  G  E  R  S  G  G  F  E  T  H  P  P  R  D  P  V  T  P  R  E  R  H  G  N
3101  CATACAAACTCGTGGCGCCGTGATCGCACCGACACCGTTCCACGTGGTAGCGGCAATACCTCAACAGACCAGGAAACGTCGCTGGGATATCCTGGATGGT   3200
       I  Q  T  R  G  A  V  I  A  P  T  P  F  H  V  V  A  A  I  P  Q  Q  T  R  K  R  R  W  D  I  L  D  G
3201  ATGGTCCGCCGCACCGTGAGCCAGAGCACTGTCGACCCAAAAACTGTGGTCATGTGCGTGTACCGTCGTGAAGAAGAAGAAACATACGACGTACTAGACG   3300
      M  V  R  R  T  V  S  Q  S  T  V  D  P  K  T  V  V  M  C  V  Y  R  R  E  E  E  E  T  Y  D  V  L  D  E
3301  AGGAGGAGCAAGACGACGACCTCCTCGGTATCCCCAACCCCACGCCACGGCGATTGAGGATATCACAAGCCGGTCCACAGCAAAACTCATGGGTGGACAC   3400
        E  E  Q  D  D  D  L  L  G  I  P  N  P  T  P  R  R  L  R  I  S  Q  A  G  P  Q  Q  N  S  W  V  D  T
3401  ACGGGGCAGGAGGGCATACGGCTCACAGGAGGAACAGGATGAGAGGACATCGACAGATCAGGTGAGTATTTTCTCCCACGACGAAACACGGGAGTTATCA   3500
       R  G  R  R  A  Y  G  S  Q  E  E  Q  D  E  R  T  S  T  D  Q  V  S  I  F  S  H  D  E  T  R  E  L  S
```

```
3501 TCACCACTGGAGTGTCCTATCGTAGGATGCACCGCGAGTTTCGTGGGCCCACGCAGATGGGAGAAGGCCAAATCCCATATATACGGGGTCCACTCGCTGG 3600
      S  P  L  E  C  P  I  V  G  C  T  A  S  F  V  G  P  R  R  W  E  K  A  K  S  H  I  Y  G  V  H  S  L  E
3601 AAGAGGTCCGCGAAATCCCGAGGGGGGAGCTTATATGTAAGGGGATAGTAAGATGCGAGACTTGTGCCACGCTCCTCCCTACGTCGGACAGAGCGAAACA 3700
      E  V  R  E  I  P  R  G  E  L  I  C  K  G  I  V  R  C  E  T  C  A  T  L  L  P  T  S  D  R  A  K  Q
3701 GGCACACCGCGACGATTGCAGACCCTATCTCCCGCGGAAAGAAAACATCCGCCGAAAAAGGGCCGCTGAAAGAGAAGCGACAGAGGCGAGCGCACAGCAA 3800
      A  H  R  D  D  C  R  P  Y  L  P  R  K  E  N  I  R  R  K  R  A  A  E  R  E  A  T  E  A  S  A  Q  Q
3801 GGAATAGCGCTACGCCTCGAGCGGCAGGGCCCGTACATAACTCCCCGCGACATAGAGGAGCCCACCAACACGACGACGGAAAGTTGGTGGAGGGAGAAGG 3900
      G  I  A  L  R  L  E  R  Q  G  P  Y  I  T  P  R  D  I  E  E  P  T  N  T  T  T  E  S  W  W  R  E  K  V
3901 TAGCTACGAAACGCTACCTTCACAGAAAGGAGTGGCCGCAGTGGCTTGACATCTGCCGCACGGTCCTCCTCGGATACTCCGCGTCATCACAAGGCGAGCG 4000
      A  T  K  R  Y  L  H  R  K  E  W  P  Q  W  L  D  I  C  R  T  V  L  L  G  Y  S  A  S  S  Q  G  E  R
4001 GCACCAACGCCAAGTGATGCTCCTTGATCTGGTCCGGAATCATCTCCACACGCGCACAGCCAGGCGCGAGCAACAGCAGCAACGTGGAAAGGATAACCAG 4100
      H  Q  R  Q  V  M  L  L  D  L  V  R  N  H  L  H  T  R  T  A  R  R  E  Q  Q  Q  Q  R  G  K  D  N  Q
4101 GAAGAGGAGGACCGCCAGAAGAAGGAGGAGAAATCCCTGCGAAACGCGTGGAAACCCTGTGCCTCCTCAGTGCGACAGGGAGGGCAGCCCAGCTCCTCGC 4200
      E  E  E  D  R  Q  K  K  E  E  K  S  L  R  N  A  W  K  P  C  A  S  S  V  R  Q  G  G  Q  P  S  S  S  Q
4201 AGCCGAAAAAGGCTCAACCGGTGGAGTACAGCCCCGAAATGGCTCAAACAATCGGGGAACTGTACCCGCAGGAGGATATCCATGATATTCCCCGGCCCACC 4300
      P  K  R  L  N  R  W  S  T  A  P  K  W  L  K  Q  S  G  N  C  T  R  R  R  I  S  M  I  F  P  G  P  P
4301 GGTGGAACAACCAGGGGTCGTGTCAGTCGACGCTGAGGAAGTAGCGAAAACTATCGCTAGGCGACTGACACGGGGCGCGGCGCCAGGGTTAGATGGGTGG 4400
      V  E  Q  P  G  V  V  S  V  D  A  E  E  V  A  K  T  I  A  R  R  L  T  R  G  A  A  P  G  L  D  G  W
4401 ACGCGAGAACTATTATACCCACTCACCCTGGACCCCGCGCTAAAGATGGAGATTGCCGCCGTTGTAAAGGACATCATAAACGCCGATGTCTCGATGGAGG 4500
      T  R  E  L  L  Y  P  L  T  L  D  P  A  L  K  M  E  I  A  A  V  V  K  D  I  I  N  A  D  V  S  M  E  V
4501 TGGGACGCCGCCTCCAAGCAACGAGCCTAACGGTACTTCGGAAGCCGAATGGGAAGTACCGACCGATTGGAGCTGAGAGCGTGTGGGCGAAGCTCGCATC 4600
      G  R  R  L  Q  A  T  S  L  T  V  L  R  K  P  N  G  K  Y  R  P  I  G  A  E  S  V  W  A  K  L  A  S
4601 CCACATAGCGATCTCCCGGGTGATGAAGACAGCCGAAAAGAAATTCTCCGGGATCCAATTCGGAGTGGGAGGCCACATCGAGGAAGCCATTGCAAAGATT 4700
      H  I  A  I  S  R  V  M  K  T  A  E  K  K  F  S  G  I  Q  F  G  V  G  G  H  I  E  E  A  I  A  K  I
4701 AGAAAAGACTTTGCAACTAAAGGCAGCCTTGCCATGCTGGATGGTCGGAACGCGTATAATGCCATCAGCAGGCGAGCCATCCTCGAGGCCGTGTACGGTG 4800
      R  K  D  F  A  T  K  G  S  L  A  M  L  D  G  R  N  A  Y  N  A  I  S  R  R  A  I  L  E  A  V  Y  G  D
4801 ACAGCACGTGGTCCCCACTATGGCGCCTCGTCAGCCTCCTCCTTGGAACCACAGGGGAGGTAGGATTCTACGAGAATGGCAAATTATGCCATACGTGGGA 4900
      S  T  W  S  P  L  W  R  L  V  S  L  L  L  G  T  T  G  E  V  G  F  Y  E  N  G  K  L  C  H  T  W  E
4901 ATCGACGAGGGGTGTAAGACAGGGGATGGTACTTGGCCCCCTGCTATTCTCCATCGGCACCTTGGCGACACTTCGCCGACTGCAGCAGACCTTCCCGGAG 5000
      S  T  R  G  V  R  Q  G  M  V  L  G  P  L  L  F  S  I  G  T  L  A  T  L  R  R  L  Q  Q  T  F  P  E
5001 GCTCAGTTTACCGCGTACCTGGACGACGTGACGGTAGCGGCACCCCCGGAAGAGCTGAAAAATGTCTGCGCAGCCACCGCTGAAGCAATGGAAGCACTCG 5100
                       4
      A  Q  F  T  A  Y  L  D  D  V  T  V  A  A  P  P  E  E  L  K  N  V  C  A  A  T  A  E  A  M  E  A  L  G
5101 GAATCGTCAACAATGCAGACAAAACCGAGGTCCTCGAACTGACTGGGGACACAGGCTTTGGGACAGCGGTGAAGCGTGTGCGCGAGTTCTTGGAGCGTAC 5200
      I  V  N  N  A  D  K  T  E  V  L  E  L  T  G  D  T  G  F  G  T  A  V  K  R  V  R  E  F  L  E  R  T
5201 GTGGCCGGATCCAATGAGCGAGGAGATTCGGGAGGGGGTGGAGAAGAAGGCGATGGAAACAGACCGCCTCTTCAAGGCAATCGTGGAGCTACCCCTCTAC 5300
      W  P  D  P  M  S  E  E  I  R  E  G  V  E  K  K  A  M  E  T  D  R  L  F  K  A  I  V  E  L  P  L  Y
5301 AACAGGACACGATGGAGGATTCTGGCGATGTCGGCAATGCCAAGGATCACATTCCTGTTGCGGAACCACGATATGCAACACACACACCGGGTGGCTTCCT 5400
      N  R  T  R  W  R  I  L  A  M  S  A  M  P  R  I  T  F  L  L  R  N  H  D  M  Q  H  T  H  R  V  A  S  W
5401 GGTTCGATGAGAGGACCACCCAGGTAATGGAGCATATTCTCGGGCAACCCATGACCGAAAGGGCCCGGAATATAGCGGCGCTGCCCGTAAGCATGGGCGG 5500
      F  D  E  R  T  T  Q  V  M  E  H  I  L  G  Q  P  M  T  E  R  A  R  N  I  A  A  L  P  V  S  M  G  G
5501 CTGTGGAATTAGGCGGATGGCCCAAGTGGCAGAGTACGCCCACCAGTGCGCCGGAGAGAAAGGTCTCCAGCAGAGGAAGACGGAGGAGGCTGACCAAAGA 5600
      C  G  I  R  R  M  A  Q  V  A  E  Y  A  H  Q  C  A  G  E  K  G  L  Q  Q  R  K  T  E  E  A  D  Q  R
5601 CAGCAAGACGACCTCTACGCCACCCTTGGGGGTGCTGATCGTCAAGTCTTTACAGCCAATACCGCCGCCGGAGCTGGCAGGCCCCTCACGGATGCTCAGG 5700
      Q  Q  D  D  L  Y  A  T  L  G  G  A  D  R  Q  V  F  T  A  N  T  A  A  G  A  G  R  P  L  T  D  A  Q  V
5701 TGAGGCTGGACGATGCCACTTTCGGAGTGTACCTGCGGGAACGTTACTGTAGGGTACTACCGGAGGGGGTCAAATGCCTATGTGGTGAAGACGCGAGCAA 5800
      R  L  D  D  A  T  F  G  V  Y  L  R  E  R  Y  C  R  V  L  P  E  G  V  K  C  L  C  G  E  D  A  S  N
5801 TCACCACATCCACACTGGCACCAAAGTGCACAATAAACCCAGGCAGATGCGACACGACATCATTAACAGCGTGTTCGCAAACGGCCTTCGCCTCTGTGGG 5900
      H  H  I  H  T  G  T  K  V  H  N  K  P  R  Q  M  R  H  D  I  I  N  S  V  F  A  N  G  L  R  L  C  G
5901 TTCCAGTGCGCGACGGAACCACGCCTAAATGAGGTGAGCAAGAGGAGGCCGGACATCCTCATTGCGGGGTTGGATACGTACGCGGTGACGGACATCACGG 6000
      F  Q  C  A  T  E  P  R  L  N  E  V  S  K  R  R  P  D  I  L  I  A  G  L  D  T  Y  A  V  T  D  I  T  V
6001 TGACGTATCCAGGGCGCGTGACCGTCGGAAACACCGCCCAAGGTCAGCGCTCAGTAGCTGCGGCAGATCCAATGAAAGCCGCATTGGTCGCGTTCCAGGA 6100
      T  Y  P  G  R  V  T  V  G  N  T  A  Q  G  Q  R  S  V  A  A  A  D  P  M  K  A  A  L  V  A  F  Q  E
6101 AAAGGAGCGCAAGTACAGCTACTGGGCGATACAAAATGGACTGGCCTTCGCACCATTTGTTATGCTTACAAACGGTGCTATTTTCGGCAAAAGTCGTGAC 6200
      K  E  R  K  Y  S  Y  W  A  I  Q  N  G  L  A  F  A  P  F  V  M  L  T  N  G  A  I  F  G  K  S  R  D
6201 TGGCTTCGCCGCGTCCTCCGGGGCCAGGACCACCGACTTACGGTAACCACCGCATTCGACGGGATAACTGCGGATGTGGTGGCAGCCGTCCTCCGCGGGA 6300
      W  L  R  R  V  L  R  G  Q  D  H  R  L  T  V  T  T  A  F  D  G  I  T  A  D  V  V  A  A  V  L  R  G  N
```

```
6301  ATGTTCACGTTTACAGTGCGGCACAAGCCCGGGGAGAGACACTTCGGTAGTTCCAGATCACTGGGATTACCAATATCCAGATGTAGAGTAGTAATAGCAA  6400
       V  H  V  Y  S  A  A  Q  A  R  G  E  T  L  R  *
6401  TAAAATAAAAACAACCCCCTGAAGAAAGGGAAGGTAATTAGCTACCAAATCATTGCCAACAGGGATCCCTCTCCACCAATCGACCGAGTAGGTCTCTTTT  6500
6501  TTCGGTTGTGCGGGCTCTCCCATAAGCCCGATGGAGAAAATCTCTTTCCATATAGGGCAATAAAATAATAATAAATAGATAGGATTATCCGGTCCATTAA  6600
6601  AGACCACGTAACCTGAAAAAGGTTACACTGCATGTTCCGTGAAAATCGGATGAGGTTTCGGAGATCAACAAAGGTGATCACGTTTAACTCGGAGGTCGGG  6700
                           4                                            1
6701  GCAGTTAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAttattagaacagtttctgtact atattggtatgagaagctcccagtaggagctgggcc  6800
6801  aacacacgcattgtgctgttggttcctgccgcatactgcgggaatctggaaggtggggtcggatgacctccactcttttttatttttttttatttttttcat  6900
6901  ttatttatttttttttgatc  6920
```

Figure 3: Nucleotide Sequence of SLACS with the predicted amino acid sequence of its ORFs. The sequence presented starts at the 5′-end of one SL-RNA coding sequence (shown in small letters). SLACS (in capitals) interrupts SL-RNA gene at base 60. The underlined 49 residues (11−60 and 6743−6792 marked 1) are the duplicated target sequences that flank the 5′and 3′junctions. The first repeat sequence at the 5′-end region is underlined and marked 2 (nucleotides 462 to 649). There is a partial SL-RNA gene sequence at the 3′-end of SLACS with its first 11 nucleotides missing (6707 to 6920). The deduced amino acid sequence of ORF1 and ORF2 are shown below the nucleotide sequence. * denotes a translation stop codon. The underlined residues in ORF1 correspond to the Cys-His motif (marked 3) and in ORF2 to the YXDD consensus sequence (marked 4).The 3′-end of SLACS contains a 36A stretch (nucleotides 6707−6742) marked 5.

Southern Hybridization Analysis

Isolation of trypanosomes, extraction of nucleic acids and Southern blot hybridization conditions have been described (17).Separation of large DNA fragments was accomplished using a PFG-unit with the hexagonal array (LKB Pulsaphor System).For the SL-RNA hybridization probe, a 1.4 kb ApaI fragment containing both the coding and flanking SL-RNA sequences was used. The fragment was isolated from Q107; ApaI ends were filled-in with Klenow; BamHI linkers were added using T4-DNA Ligase, and it was subcloned into BamHI cleaved M13mp18 vector DNA. Construction of SLACS probe, pSB1 has been described (17). Replicative form m13 recombinant DNAs were labelled in hybridization experiments using the random primer kit (Boehringer Mannheim) with α^{32}P-dATP.

Subcloning and Sequencing Strategy

For determining the DNA sequence information, restriction enzyme fragments of the recombinant phage Q107 were cloned directly into the M13mp18 and M13mp19 sequencing vectors (21). The specific restriction enzyme sites that were used and the direction of sequence determined are shown in Figure 2B. Single-stranded template DNA was prepared (22) and the nucleotide sequence was determined using the dideoxy chain termination method modified by the use of α[^{35}S]dATP(23).

In addition to cloned restriction enzyme fragments, a deletion subcloning approach was also used (24). Single-stranded DNA from the M13 recombinant clones containing the 4.5 kb BglII-BamHI and the 2.2 kb SalI-EcoRI fragments was used to generate deletion clones with the Cyclone kit obtained from International Biotechnologies Inc. Where indicated in Figure 2B, synthesized oligonucleotide probes were used as sequencing primers.

RESULTS

Genomic organization of the SLACS element

Our previous pulse field gel electrophoresis analysis has shown that all SL-RNA genes are tandemly organized into two large fragments (17 and Figure 1A). These fragments are generated by restriction enzymes that neither cleave within SL-RNA repeat sequences nor within SLACS retrotransposons; i.e. EcoRI. When DNA is digested with restriction enzymes which cleave SLACS but not SL sequences, nine fragments hybridize to SL probes (Figure 1B,Lane 1). When the same genomic digests are probed with DNA fragments containing only SLACS sequences, the hybridization pattern is found to be identical, i.e. the same two large (Figure 1C,Lane 1) and the same nine smaller fragments (Figure 1B,Lane 2 and Figure 1C Lanes 2,3) are detected. Thus there are nine SLACS sequences, and all are located within the SL-RNA gene clusters.

Nucleotide sequence of a SLACS element

The organization of a SLACS element inserted within one SL-RNA gene sequence is schematically shown in Figure 2A. The cloning strategy used for obtaining the primary sequence is summarized in Figure 2B. The complete nucleotide sequence of the SLACS retrotransposon is shown in Figure 3.

SLACS retrotransposon is 6678 bp long. The DNA sequence presented in Figure 3 starts at nucleotide 1 at the 5′-end of the SL-RNA gene coding region. SLACS sequences interrupt the SL gene at nucleotide 60. The 3′-end of the element has a stretch of 36 A residues. A partial SL-RNA gene flanks the 3′-end border. This partial gene starts at nucleotide 11 of the SL coding sequence and extends into the repeat unit sequences. The recombinant Q107 which carries SLACS sequences has four tandemly organized 1.4 kb SL-RNA gene units at the 3′-end of the retrotransposon (17). There has been a 49 bp target DNA duplication at the site of SLACS insertion and this duplication corresponds to nucleotides 11 to 60 of the SL-RNA gene sequence. There is no repeated sequence at both ends of SLACS in either direct or inverted orientation.

Our Southern blot analysis had shown that there are no truncated SLACS sequences in the *T.b.gambiense* genome and that all nine copies of SLACS have the same overall organization (17). However, we noted that the elements varied from 6 to 7.2 kb in size.

There is a repeated sequence in the 5′-end variable region

The variability at the 5′ half of different copies of SLACS was located within a PvuII restriction enzyme fragment (17). The DNA sequence of this region reveals a repeated segment of about 185 bp corresponding to nucleotides 462 to 1173. The SLACS element sequenced from Q107 has three complete repeat segments of 187, 184 and 185 bp length. The sequence comparison of the repeats shows that several base pair insertions and deletions have accumulated. The fourth repeat has homology with the initial 154 nucleotides.

Protein coding sequences of SLACS

We have determined the positions of methionine start and chain-terminating stop codons in each of the six reading frames of

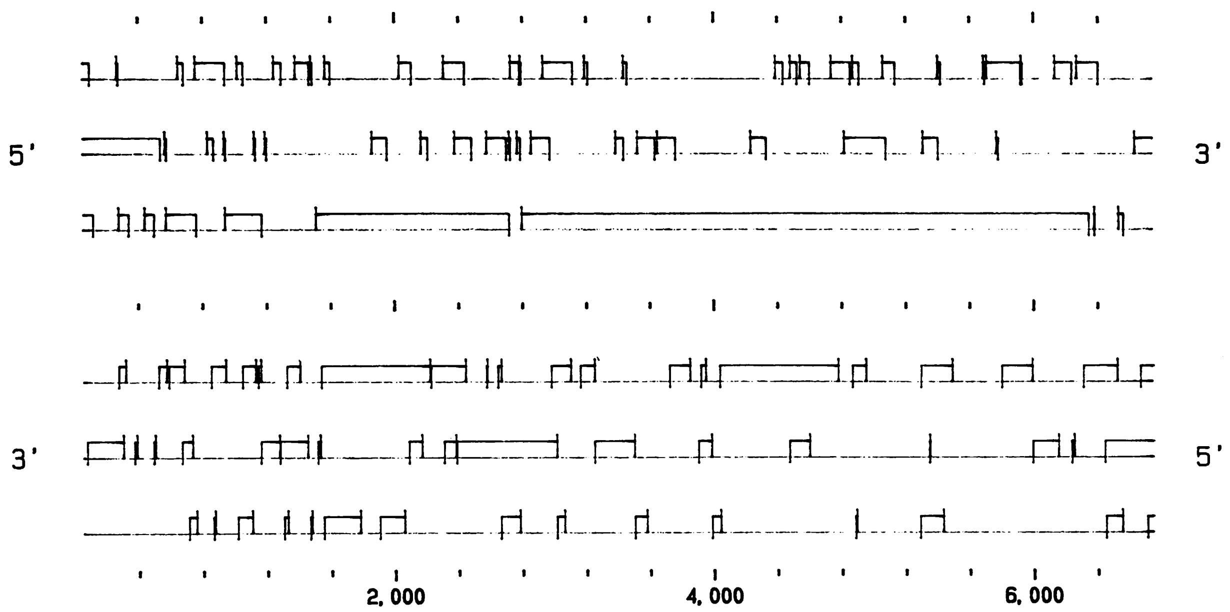

Figure 4: Protein Coding Sequences of SLACS. Potential protein coding sequences of SLACS are boxed in the six reading frames. Only the start and stop codons are shown for each open reading frame (ORF). Both ORF1 and ORF2 are in Frame 3 in the 5'to 3'direction and are separated by 79 nucleotides.

SLACS sequence (Figure 4). Reading frame 2 contains an ORF of 154 amino acids at the 5'-end of SLACS upto nucleotide 463. None of the frames contain any potential protein coding sequences between nucleotides 463 to 1177. This region corresponds to the variable region containing the 185 bp repeats. Starting at nucleotide 1512, SLACS contains two extensive open reading frames (ORFs), both in reading frame 3, together occupying 75% of its length. ORF1 codes for 384 amino acids, starting at position 1512 to 2726. ORF2 is 1182 amino acids and starts with the AUG methionine codon at position 2802 and extends to nucleotide 6530. The ORFs are separated by 79 nucleotides. Upstream of ORF1 the sequence between nucleotides 1180 to 1512 in Frame 3 contains no stop codons. Thus ORF1 could potentially be longer by 110 amino acids, however, the first methionine start codon is found at position 1512.

ORF1 of SLACS exhibits DNA Binding Properties

The primary translation product of the gag gene in retroviruses is a polyprotein that is subsequently cleaved to generate virion core polypeptides. One of these polypeptides is the highly basic nucleic acid binding protein that originates from a domain in the 3'-terminal portion of the gene. Nearly all retroviral protein sequences contain either one or two groups of conserved amino-acids in this region called Cys motifs (25). Non-LTR containing retrotransposons have similar cys-motifs associated with the 3' portions of their ORF1 and this conserved is generally $Cysx_2Cysx_4Hisx_4Cys$ where x may be any amino acid (3). However a somewhat different spacing of Cys and His residues has also been found in a number of regulatory eukaryotic DNA-binding proteins (14). The ORF from the transposable element TRS-1 also codes for DNA binding sequence motifs. This element is also characterized in the *T. brucei* genome, but its primary sequence is different from SLACS and it represents a highly repeated sequence distributed randomly in the genome (13). A CysHis motif is found to be repeated five times in the last third of the putative TRS-1 ORF but with a spacing that is different from that of other non-LTR retrotransposons,ie. $Cysx_2Cysx_{13}Hisx_5His$ (26). The ORF1 in SLACS also has one CysHis motif with identical spacing as the modified motif in TRS-1 (Figure 3). This sequence in SLACS is present at the 3'-end of ORF1. A similar arrangement has been found in the non-LTR retrotransposons.

ORF2 may code for a reverse-transcriptase-like enzyme

By comparing the retroviral pol genes and the ORF of the Drosophila element 17.6, Toh et al. identified a 175 amino acid stretch in which 33 positions were either conserved or had functionally equivalent amino acids (27,28). Recently Xiong and Eickbush compared the amino acid sequences of the non-LTR retrotransposons in this domain and found that in 8 regions non-LTR elements share a greater similarity to each other than to the retroviruses or Copia-like elements (3). These conserved residues are also coded by the ORF2 of SLACS. A comparison of SLACS ORF2 coded amino acid residues and those coded by the other non-LTR retrotransposons is shown in Figure 5. In a comparison of the eight homologous regions, SLACS has 11 of the 12 invariant amino acid residues identified by Toh et al. for all reverse-transcriptase containing sequences. A Proline in region 5, the tyrosine residue in the YXDD box in region 6 and a glycine in region 7 are conserved both in SLACS and in LTR-containing elements. However, they are found to vary in the non-LTR retrotransposons (3). When compared only with non-LTR retrotransposons, SLACS has 24 of the 32 residues that are found to be invariant by Xiong and Eickbush (Figure 5). In addition, 2 of the conserved residues in region 4 contain chemically similar amino acids in SLACS. Among the residues that were identified as similar in non-LTR sequences in region 5, five have conservative amino acid substitutions in SLACS at these positions. Furthermore, the spacing between the eight conserved regions is in agreement with those found for the non-LTR elements. The YXDD consensus box in region 6 that is conserved in all reverse-transcriptase-like enzymes, contains an alanine residue for X in all of the non-LTR retroposons. SLACS, however, has a leucine residue in this position. This arrangement is similar in the

```
                          1                                      2                        3
                        -  -- -                           -     ---     -       -        - -
SLACS..(530)..RLTRGAAPGLDGWTRELLYP..(35)..LTVLRKPNG__K_____YRPIGAESVWAKLASHIAISR..(12)..QFGVGGHIEE

 Ingi..(145)..LLPSGSAAGPDCLYNEALQH..(38)..IIPILKAGKKAE_DLDSYRPVTLTSCLCKVMERILAAR..(12)..QSGFRPGCST
   R1..(461)..SLK_WTAPGIDGLTARIIKK..(38)..LLVLPKGNGRPLTDPKAYRPVTLLPVLGKILEKVLLQC..(11)..QKGFSPGRST
   R2..(436)..RFWR_TSPGPDGIRSGQWRA..(35)..TVFVPKVERP__GGPGEYRPISIASIPLRHFHSILARR..(10)..QRGFICADGT
 L1Md..(499)..SLPTKKSPGPDGFSAEFYQT..(38)..ITLIPKPQKD_PTKIENFRPISLMNIDAKILNKILANR..(12)..QVPFIPGMQG
 L1Hs..(489)..SLPTKKSPGPDGFTAEFYQR..(30)..IILIPKPGRDTTKKEN_FRPISLMNIDAKILNKILANQ..(12)..QVGFIPGMQG
    I..(310)..TLK_GCAPGLNRISYQMIKN..(38)..IIPILKAGKK_AEDLDSYRPVTLTSCLCKVMERILAAR..(14)..QFGFKKGKST
    F..(447)..NLSPKKSPGYDLITPEMIIQ..(38)..IIMIPKPGKN_HTVASSYRPISLLSCISKLFEKCLLIR..(14)..QFGFRESHGT
                L     PG D                    K              YRP  L     K     L  R          Q GF
                                                              R

                     4                     5                     6               7            8
                  --   -▲▲          -  --  ▲ * ▲▲ ▲    -         * --            *      -       --
SLACS..(17)..LDGRNAYNAISR..(46)..GVRQGMVLGPLLFSIGTLATL..(13)..AYLDDVTVAA..(19)..GIVNNADK..(36)..EGV

 Ingi..(23)..VDYEKAFDTVDH..(45)..GVPQGTVPGSIMFIIVMN_SL..(12)..FFADDLTLLA..(26)..SVNVAKTK..(29)..LGV
   R1..(23)..LDISGAFDNAWW..(42)..GCPQGSVLGPTLWNVLMDDLL..(12)..AYADDVTVLV..(20)..GWA.SRNK..(38)..LGV
   R2..(25)..LDFAKAFDTVSH..(45)..GVRQGDPLSPILFNYVMDLIL..(19)..AYADDLVLLA..(20)..GLRLNCRK..(41)..LGV
 L1Md..(26)..LDAEKAFDKIQH..(45)..GTRQGCPLSPYLFNIVLE_VL..(23)..LLADDMIVYI..(21)..GYKINSNK..(31)..LGV
 L1Hs..(26)..IDAEKAFDKIQQ..(45)..GTRQGCPLSPLLFNIVLE_VL..(23)..LFADDMIVYL..(21)..GYKINVQK..(31)..LGI
    I..(22)..LDFSRAFDRVGV..(45)..GIPQGSPISVILFFIAFN_KL..(13)..AYADDFFLII..(26)..GASLSLSK..(31)..LGR
    F..(25)..LDVSQAFDKVWL..(44)..GVPQGSVLGPTLYLIYTA_DI..(07)..TFADDTAILS..(26)..RIKVNEQK..(32)..LGV
             LD   AFD            G  QG      L      L           ADD                       K          LGV
              D    F                P G      P                 Y DD               G      K           G
```

Figure 5: Comparison of ORF2 sequences in the reverse-transcriptase coding domain. SLACS ORF2 and the putative reverse transcriptase coding ORFs of other non-LTR retrotransposons are aligned for maximum homology. Conserved amino acid positions are grouped into eight regions as has been aligned by Xiong and Eickbush (3). The first number in parenthesis indicates the distance of the putative homologous region from the beginning of its ORF; subsequent numbers show the amino acids deleted from the sequence for alignment. The bars (−) above the residues show invariant positions that are conserved in SLACS while the triangles (▲) show the residues that have conservative substitutions. The stars (*) denote residues identified by Toh et.al. which are conserved in SLACS but not necessarily present in all non-LTR retrotransposons. The first row of letters at the bottom indicate conserved residues shared by non-LTR retrotransposons identified by Xiong and Eickbush (3). The second row of letters in bold show the residues in these eight regions that are identified by Toh (27,28). The sequences analyzed were taken from : INGI (11), R1 (3), R2 (4), L1Md (10), L1Hs (9), I (6) and F (5).

retroviral sequences. We have also noted that a second TAYLDD sequence is coded for by nucleotides 5605−5623 in the same reading frame as ORF2 but it is in the 3′ to 5′ direction.

DISCUSSION

The DNA sequence analysis of SLACS indicates that it belongs to the group of non-LTR containing insertion elements. It has generated a 49 bp target DNA duplication at its insertion site; its 3′-end contains an extensive poly(A) stretch and it has two ORFs spanning more than 75% of its sequence. Its ORF1 contains a CysHis motif which has been associated with metal binding domains in nucleic acid binding proteins whereas the longer ORF2 sequence shows homology with reverse-transcriptase coding elements. An unusual feature of SLACS is that it is present only in 9 copies and is associated only with the SL-RNA genes. Furthermore, the organization of all nine copies of SLACS is conserved; that is there are no truncated copies of SLACS present in *T.b.gambiense*. Typically, Line-1 like retrotransposons are dispersed throughout the genome in high copy numbers and many truncated versions are present. These unusual findings in SLACS might suggest that SLACS insertion represents a target site specific event that may be of recent evolutionary origin.

All nine copies of SLACS could either represent independent insertion events or alternatively may be due to one event that has subsequently been duplicated by a recombination mechanism such as unequal sister-chromatid exchange. Thus, to investigate this possibility, we have obtained the 5′ and 3′ junction sequences from a second SLACS element (17). Differences in the reduplicated target site DNA that flank both ends of this group of retrotransposons might be one reason to postulate insertion events due to independent transposition mechanisms. We found however, that the second SLACS element had inserted into the same coding region of one SL-RNA gene and had the same 49 bp target site duplication flanking both of its ends (17). It contained a stretch of only 12 A residues at its 3′-end and by restriction mapping analysis. This element had six copies of the 185 bp repeat sequence located 5′ to ORF1 instead of the three repeats found in the SLACS element described here. Based on target site duplication sequences, we cannot rule out the possibility that an unequal sister chromatid exchange mechanism has led to multiple SLACS copies but the variation in the 185 bp repeats and in the poly(A) stretch makes this explanation less likely.

Transcription initiation signals in the non-LTR retrotransposons are thought to be present at the 5′-end of the elements. The 5′-end 200 bp segment of the drosophila element, Jockey has been shown to contain promoter activity (31). The 5′-end of the L1 element in the mouse also contains transcription initiation signals (Diana Severynse, personal communication). The organization of the mouse L1 elements shows that there is a repeated sequence of 200 bp length located 5′ of ORF1. The number of these repeats varies in the different L1 sequences in the mouse, similar to our findings in SLACS. These 200 bp repeats in the L1 element are shown to contain promoter activity in gene fusion constructs (Diana Severynse, personal communication). It remains to be shown whether the 185 bp repeated sequence at the 5′ region of SLACS might also function as transcription initiation sites.

In *C.fasciculata*, all copies of the related element CRE-1 have also been inserted into the same target site within the SL-RNA coding sequence; i.e. between nucleotides 11 and 12. The element

CRE-1 is flanked by target DNA duplication at its site of insertion and lacks LTRs. Both SLACS and CRE-1 contain a hydrophobic residue for X in their YXDD box. In this respect, CRE-1 and SLACS resemble the retroviruses and LTR-containing elements since all other non-LTR retrotransposons have alanine at this position. Our preliminary observations suggest that SL-RNA interrupting sequences are also present in a new world trypanosome, *T.cruzi*. Similar SL-RNA interrupting sequences are also reported in another trypanosomatid, *Leptomonas seymorii* (29). How closely these sequences are related to SLACS remains to be shown. Either the retrotransposons in different species might represent recent insertion events that have a common target site insertion specificity, or they may have evolved from a common precursor. There is a precedent for a family of evolutionarily related elements inserted into a conserved site within the insect ribosomal genes. Eickbush has found that the retrotransposons R1 and R2 are present in a wide variety of insect species and have maintained their target site specificity but at the same time have diverged independently in their individual sequences (Thomas Eickbush, personal communication). The SL-RNA gene associated retrotransposon sequences might also represent a similar group of elements in the members of the family Trypanosomatidae.

Three other mobile elements have been characterized in the African trypanosome; RIME (30), INGI/TRS-1 (11,12) and MEA (18). Based on its genomic organization and limited sequence comparison MEA, most likely represents the same sequence as SLACS. RIME is present in many copies and also resulted in a 7 bp duplication of target sequences flanking its both ends. It has an ORF encoding a potential protein of 160 amino acids and the 3′-end of the element is preceeded with a stretch of 14A residues. INGI or TRS-1 are longer dispersed highly repetitive elements associated with RIME sequences at both ends. The ORF of INGIor TRS-1 has both the invariant residues conserved in reverse-transcriptase-like sequences and five copies of the CysHis motif associated with the gag polypeptides. A comparison of the reverse-transcriptase domain in ORF of SLACS and INGI does not show any greater similarity between these two elements than are found in the other LINE-1 like elements. However, the unusual CysHis pattern in the gag polypeptides is conserved in both retrotransposons. This may suggest either an evolutionary or a functional relatedness for this domain.

It remains to be shown what the functional significance of the SLACS is. We do not detect any RNA transcripts corresponding to these sequences. This could be a consequence of either low abundance of the RNA product(s) or instability associated with these transcripts. Alternatively, the expression of SLACS might be under developmental regulation which permits function only in specific stages of the parasite life-cycle.

ACKNOWLEDGEMENT

We thank Drs. Peter Mason, Ben Beard and Ditas Villenuava for their critical reading of the manuscript. We thank Dr.Russell Doolittle for his help in the comparative sequence analysis. We thank Dr.Martine Armstrong for her help and Louise Camera-Benson for her technical assistance. This work was supported by a grant to the Yale Center for Molecular Parasitology by the MacArthur Foundation and also by a grant from NIH (NIH AI08614). S.A. and F.F.R. are Investigators of the MacArthur Foundation Consortium on the Biology of Parasitic Diseases.

REFERENCES

1. Singer,M.F. (1982) Cell 28:433–434
2. Singer,M.F. and Skowronski,J. (1985) Trends Biochem.Sci. 10:119–122
3. Xiong,Y. and Eickbush, T.H. (1988) Mol.Cell Biol. 8:114–123
4. Burke,W.D., Calalang,C.C. and Eickbush,T.H. (1987) Mol. Cell. Biol. 7:2221–2230
5. DiNocera,P.P. and Casari,G. (1987) Proc.Natl.Acad.Sci.USA 84:5843–5847
6. Fawcett,D.H., Lister,C.K., Kellett,E. and Finnegan,D.J.(1986) Cell 47:1007–1015
7. Bucheton,A., Paro,R., Sang,H., Pelisson,A. and Finnegan,D.J. (1984) Cell 38:153–163
8. Dawid,I.B., Long,E.O., DiNocera,P.P. and Pardue,M.L. (1981) Cell 25:399–408
9. Scott,A.F., Schmeckpeper,B.J., Abdelrazik,M., Comey,C.T., O'Hara,B., Rossiter,J.P., Cooley,T., Heath,P., Smith,K.D. and Margolet,L. (1987) Genomics 1:113–125
10. Loeb, D.D., Padgett,R.W., Hardies,S.C., Shehee,M.B., Comer, M.B., Edgell, M.H. and Hutchison,C.A. (1986) Mol.Cell.Biol. 6:168–182
11. Kimmel,B.E., Ole-Moiyoi,O.E. and Young,J.R. (1987) Mol.Cell.Biol. 7:1465–1475
12. Murphy,N.B., Pays,A., Tebabi,P., Coquelet,H., Gyalex,M., Steinert,M. and Pays,E. (1987) J.Mol.Biol. 195:855–872
13. Doolittle,R.J., Feng,D.F., Johnson,M.S., McClure,M.A. (1989) Q.Rev.Biol. In press.
14. Xiong,Y. and Eickbush,T.H. (1988) Mol.Biol.Evol. 5(6): 675–690
15. Berg,J.M. (1986) Science 232:485–487
16. Xiong,Y. and Eickbush, T.H. (1988) Cell 55:235–246
17. Aksoy,S., Lalor,T.M., Martin,J., Van der Ploeg,L. and Richards,F.F. (1987) EMBO 6:3819–3826
18. Carrington,M., Roditi,I. and Williams,R.O. (1987) Nucl.Acids.Res. 15:10179–10198
19. Boeke,J.D. and Corces,V.G. (1989) Ann.Rev.Microbiology 43:403–434
20. Merritt,S.C., Tschudi,C., Konigsberg,W.H. and Richards,F.F (1983) Proc.Natl.Acad.Sci. USA, 80:1536–1540
21. Messing,J. (1983) Methods Enzymol. 101:28–78
22. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc.Natl.Acad.Sci.USA, 74:5463–5467
23. Sanger,F. Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) J.Mol.Biol. 143:161–178
24. Dale,R.M.K., McClure,B.A. and Houckins,J.P. (1985) Plasmid 13:31–40
25. Covey, S.N. (1986) Nucl. Acid Res. 14: 623–633
26. Pays,E. and Murphy, N.B. (1987) J.Mol.Biol. 197:147–148
27. Toh,H.R., Kikuno,R., Hayashida,T., Miyata,T., Kugimiya,W., Inouye,S., Yuki,S. and Saigo,K. (1985) EMBO 4:1267–1272
28. Toh, H., Hayashida,H. and Miyata,T. (1983) Nature (London) 305:827–829
29. Bellofatto,V., Cooper,R. and Cross,G.A.M. (1988) Nucl.Acids.Res. 16:7437–7456
30. Hasan,G,, Turner, and Cordingley,M.J. (1984) Cell 37:333–341
31. Mizrokhi, L.J., Georgieva, S.G. and Ilyin, Y.V. (1988) Cell 54: 685–691