The EMBO Journal vol.6 no.12 pp.3819–3826, 1987

Multiple copies of a retroposon interrupt spliced leader RNA genes in the African trypanosome, *Trypanosoma gambiense*

Serap Aksoy, Thomas M.Lalor[1], Jennifer Martin[2], Lex H.T.Van der Ploeg[3] and Frank F.Richards

Yale MacArthur Center for Molecular Parasitology, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06510 and [3]Department of Genetics and Development, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA
[1]Present address: Pathology Department, Tufts Medical School, 136 Harrison Ave., Boston, MA 02111, USA
[2]Present address: Anatomy and Cell Biology Department, Tufts Medical School, 136 Harrison Ave., Boston, MA 02111, USA

Communicated by P.Borst

The 140-nucleotide spliced leader (SL) RNA, involved in mRNA maturation in the African trypanosomes and in other kinetoplastida, is encoded by a tandem array of spliced leader genes. We show that the 1.4-kb SL gene repeat unit in *Trypanosoma gambiense* is organized in tandem arrays confined to two large (minimum size 350–450 kb) restriction fragments. SL genes in both arrays are interrupted by a total of eight conserved insertion elements. Cleavage of genomic DNA at restriction sites present within the insertion element but not in the SL gene repeat, releases variable numbers of SL genes from the tandem array. Since the insertion element contains a terminal poly(A) track of 36 bases and because a 49-bp duplication of target DNA has occurred at the integration site, we conclude that it is a retroposon. This retroposon is uniquely associated with the SL gene clusters. These retroposons presumably originated from a single insertion event after which their copy number increased, possibly through unequal sister chromatid exchange.

Key words: orphon/retroposon/spliced leader RNA/trypanosome

Introduction

mRNA maturation in African trypanosomes and other kinetoplastida involves a splicing mechanism which is different from that of most other eukaryotes. All mature mRNA consist of two exons that are discontinuously transcribed from separate genes and joined at a later stage. A common non-coding 35-nucleotide leader is found at the 5′ end of all mRNAs studied. This leader is encoded on a small 140-nucleotide RNA referred to as the spliced leader (SL) RNA (Campbell *et al.*, 1984; Kooter *et al.*, 1984; Milhausen *et al.*, 1984). Several different models have been presented to explain discontinuous transcription. It has been proposed that transcription initiation requires SL RNA as a primer joining the separate transcripts; or that alternatively, *trans*-splicing of SL and main coding exons occurs. Recently, evidence has been presented that supports the *trans*-splicing model (for a review, see Borst, 1986; Van der Ploeg, 1986).

In *Trypanosoma brucei*, the SL RNA is encoded by ~200 genes, which are organized as tandem arrays of a directly repeated 1.4-kb unit (DeLange *et al.*, 1983; Nelson *et al.*, 1983). When genomic DNA is digested with certain restriction enzymes that do not cleave within the repeat unit, high mol. wt fragments are produced that contain many SL gene repeat units. However, small fragments containing the SL gene are also found. These small fragments have been referred to as the SL gene orphons and these fragments vary in size and number among different *T. brucei* isolates (Nelson *et al.*, 1983; Parsons *et al.*, 1984, 1986).

We show here that all spliced leader genes are clustered on two large restriction fragments (at least 350 and 450 kb in size). The tandemly repeated SL genes of both clusters are interrupted at multiple sites by a conserved sequence of ~7 kb. When these sequences are cut by enzymes not found in the SL monomer, one observes the appearance of orphon restriction enzyme fragments containing the SL genes which have been released from the tandem array. The conserved sequence is associated only with

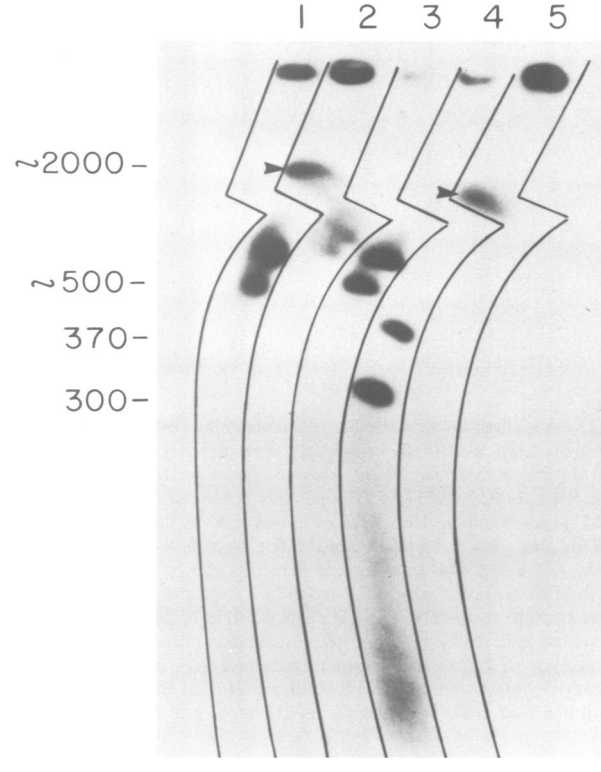

Fig. 1. Pulsed field gel analysis of *T.gambiense* restricted chromosomes hybridized to SL gene probe. *T.gambiense* chromosomes embedded in 0.5% agarose blocks were cleaved with restriction enzymes (see Materials and methods for further details) *Sfi*I, *Not*I, *Xba*I and *Hin*dIII in **lanes 1–4** respectively, **lane 5** contains intact chromosomes from a heterogenous population of *T.brucei* trypanosomes obtained from a parasitemic rabbit at day 28 after infection, d28, and was used as the size marker (Shea *et al.*, 1986). A pulse time of 40 s was used at 210 V and electrophoresis was carried out at 12°C for 60 h. Following acid hydrolysis (0.25 M HCl for 15 min), denaturation and neutralization, DNA was transferred to nitrocellulose membrane according to the Southern protocol (Southern, 1975). Probe pSLpv5 was used for hybridization according to the conditions described in Materials and methods. A 2-h exposure is shown with no intensifying screen. *Not*I enzyme analysis in **lane 2** shows a partial cleavage pattern. On subsequent complete digests, no homology with the ~2-Mb fragment is seen with this enzyme. The arrows indicate the 2-Mb fragments.

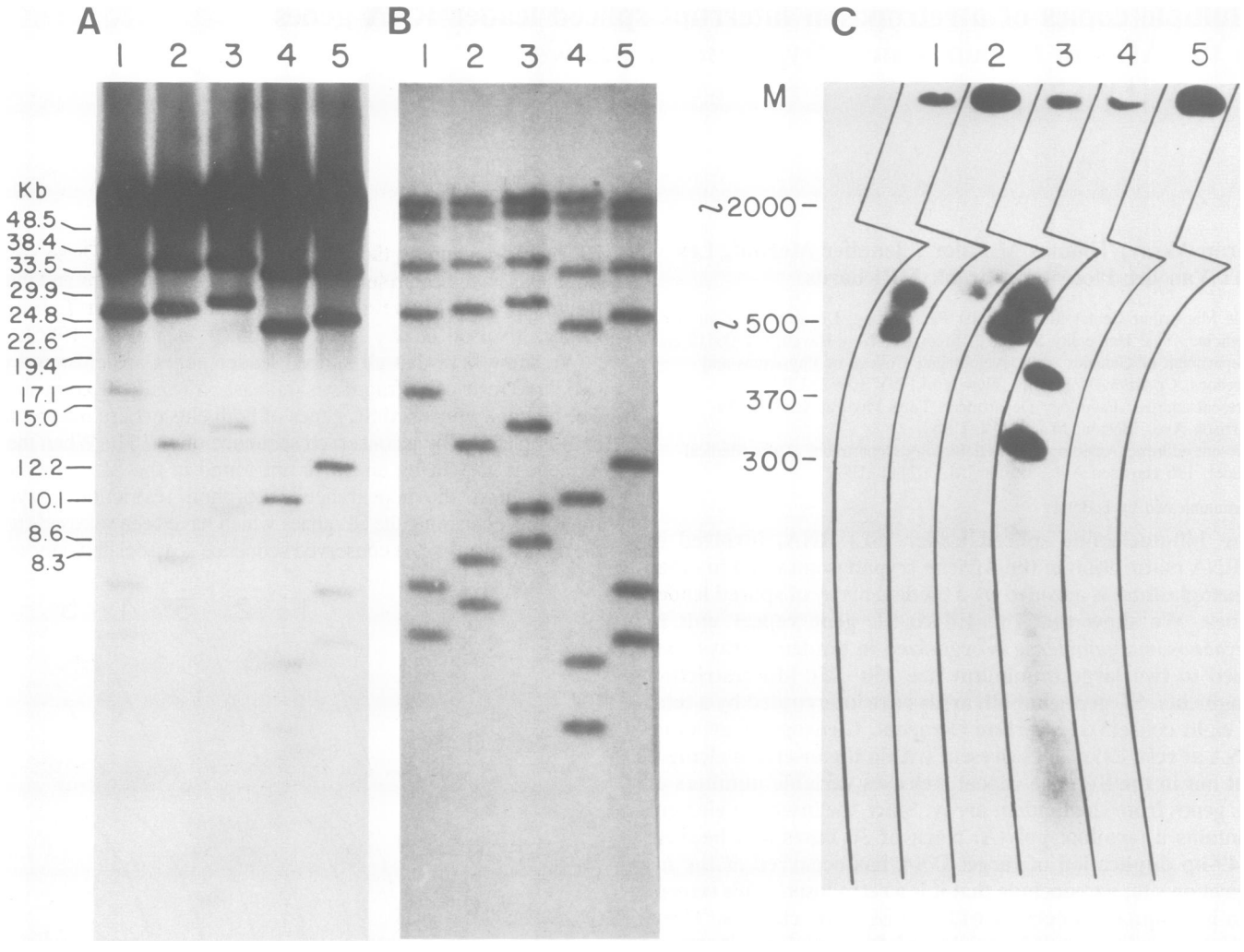

Fig. 2. Autoradiogram showing SL gene-containing restriction fragments. (**A**) SL gene homologous fragments generated by the restriction enzymes *Bam*HI, *Sal*I, *Bgl*II, *Sma*I and *Hind*II in **lanes 1–5** respectively. **Lane M** has the high mol. wt lambda DNA size standard co-electrophoresed as the size marker (Bethesda Research Laboratories). Digested genomic DNA (2 μg) was analyzed by gel electrophoresis overnight on 0.4% agarose gels in TAE buffer (0.04 M Tris acetate, 0.002 M EDTA). The gel was denatured, neutralized and subsequently transferred to nitrocellulose by the Southern protocol (Southern, 1981). The SL gene-containing clone, pSLpv5, was used as the hybridization probe in Figure 3A. Hybridization conditions were 50% formamide, 5 × SSC, 5 × Denhardt's and 0.2% SDS at 42°C for 36 h. 5 × 10^5 c.p.m./ml of the hybridization probe was used and the filters were washed at 65°C in 0.2 × SSC for 3 h. A 4-h exposure time with no screen is shown. (**B**) The digestion, electrophoresis and hybridization conditions are identical to those in (**A**) except that the hybridization probe was pSB1 and a 24-h exposure time is shown. The same digests were also analyzed by PFG in order to size the high mol. wt fragments more accurately. By using a pulse time of 20 s for 18 h followed by a pulse time of 5 s for an additional 18 h, we were able to show that the high mol. wt bands could be resolved into three bands of ~100, ~150 and ~175 kb all of which contain pSB1 and pSLpv5 homology (data not shown). (**C**) The PFG analysis of *T.gambiense* chromosomes prepared, cleaved and analyzed by conditions described in Figure 1, hybridized to SLACS DNA probe, pSB1. The enzymes used were *Sfi*I, *Not*I, *Xba*I and *Hind*III in **lanes 1–4** respectively. An overnight exposure with an intensifying screen is shown. Hybridization in the *Sfi*I and *Not*I wells is due to partial cleavage since *Xba*I and *Hind*III digests have minor amounts of DNA trapped at the well.

the SL gene cluster(s) and has many of the characteristics of a retroposon. The evolution of the repertoire of spliced leader genes and the presence of retroposons is discussed.

Results

All SL genes are clustered on two large chromosomal DNA fragments

When *T. brucei* DNA is cleaved with restriction enzymes such as *Sfi*I, *Xba*I or *Hind*II, SL genes are only found in high mol. wt DNA (Nelson *et al.*, 1983). We determined the organization of the SL gene locus by separating the DNA fragments with pulsed field gel electrophoresis (PFG) (Schwartz *et al.*, 1983; Schwartz and Cantor, 1984). Analysis of restriction enzyme-digested *Trypanosoma gambiense* chromosomes reveals two large fragments that show SL gene homology when separated by PFG (Figure 1). Digestion with *Sfi*I generates fragments of ~500 and 600 kb (lane 1), whereas with *Not*I two larger fragments of ~600 and 700 kb observed (lane 2). *Xba*I also generates two fragments of 500 and 600 kb that hybridize with the SL gene probe (lane 3). Similar analysis with *Hind*III reveals fragments of 350 and 450 kb (lane 4). In intact trypanosome chromosomes, SL gene homology is located in the well and in the 2-Mb fragment(s) (lane 5). Hybridization observed in the wells in *Sfi*I and *Not*I digests is due to incomplete cleavage since in *Xba*I and *Hind*III digests in lanes 3 and 4, respectively, we do not detect any major hybridization in the wells. This suggests that all SL genes are restricted to two chromosomal sites.

However, some other restriction enzymes which do not cut the SL gene repeat can generate a constant number of low mol.

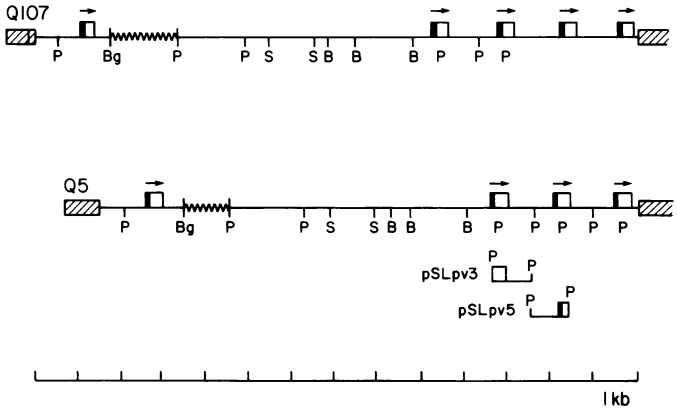

Fig. 3. Restriction maps of recombinant phages containing SL gene repeat units and its flanking DNA sequences. The hatched regions denote the phage arms; Q107 contains a 14-kb *T.gambiense* DNA insert while that of Q5 is 13 kb. The boxed regions show the 140-nucleotide coding segment of the SL gene repeat units with the 5′ end 35-mer sequence darkened; the arrows above indicate the orientation of the SL RNA transcription. Q107 contains four SL gene repeat units while Q5 has only three at the 3′ end of the insert. Both recombinants have SL gene homology at their 5′ end. Some of the relevant restriction sites in the 7-kb SL flanking DNA are noted. The 1.2-kb *BglII*−*PvuII* fragment in Q107 and the corresponding 1-kb *BglII*−*PvuII* fragment in Q5 are shown by zig-zag lines. Restriction sites: B, *Bam*HI; Bg, *BglII*; P, *PvuII*; and S, *SalI*.

wt SL fragments. Analysis of these digests on low-percentage agarose gels always generates eight SL homologous bands which are of different lengths, depending on the enzyme used (Figure 2A). To study the genomic environment of the SL sequences present on these fragments, we have isolated recombinant phages with SL gene-containing inserts.

Characterization of recombinant phages carrying SL genes

A *T. gambiense* genomic DNA library (Tschudi *et al.*, 1985) was screened with clone pSLpv5, a 650-bp *PvuII* fragment that carries 60 bases from the 5′ end of the SL gene coding segment and a contiguous 590-base segment of the repeat unit sequence (Figure 4). The restriction enzyme map of two SL-positive phages is shown in Figure 3. Phage Q107 is found to contain, at its 3′ end, four 1.4-kb SL gene repeat units arranged in tandem. The 7-kb segment adjacent to the 5′ end of the last tandem SL gene does not resemble the SL genes in its restriction map and does not hybridize either to the SL gene or to repeat unit probes. However, at the 5′ end of the Q107 insert, there is another SL homologous region adjacent to the 7-kb non-SL sequence. Phage Q5 shows the same overall organization with only three tandemly repeated SL genes located at the 3′ end of the 7-kb non-SL sequence. The 7-kb DNA segment adjacent to the SL gene repeat units in both phages are similar in their restriction enzyme pattern. There is only one restriction enzyme fragment (*PvuII*−*BglII*) within this 7-kb non-SL sequence that shows a variation in size. Thus, both recombinant phages carry a 7-kb non-SL homologous sequence flanked by spliced leader genes at both ends. We refer to this 7-kb SL-associated conserved sequence as SLACS.

Every SL-containing DNA fragment has SLACS homology

We wanted to see if the 7-kb SLACS DNA associated with SL sequences in two of our recombinant phages is a conserved feature of all SL gene orphons. A 216-bp *Bam*HI−*Sau*3A fragment located at the 3′ end of the 7-kb sequence in Q107 was isolated and cloned into the M13mp8 vector (Figure 3). This clone (pSB1) shows no homology with either the SL gene or the repeat unit sequence (data not shown). When pSB1 is used as the hybridization probe on genomic DNA digested to reveal orphon fragments, eight bands of equal intensity are resolved (Figure 2B). These bands correspond to the multiple SL gene homologous fragments observed before (Figure 2A). Thus, every SL-containing fragment appears to be associated with at least part of the 7-kb SLACS DNA. We determined that the copy number of the pSB1 clone is 8−12 per diploid genome, which is compatible with our Southern blot analysis from which eight fragments are predicted (data not shown).

When probe pSB1 is hybridized to trypanosome chromosomes digested to reveal the length of the complete SL cluster(s), and the DNA separated by PFG (as described in Figure 1), the same two high mol. wt fragments that hybridize to the SL gene probes are again observed (Figure 2C). Thus, both high mol. wt restriction fragments carry all of the SLACS segments, as well as the tandem arrays of SL repeats.

The entire 7-kb SLACS DNA is associated with every SL gene-containing fragment

Because a detailed restriction map of the SLACS DNA present on phages Q107 and Q5 was known, it was possible to determine the arrangement of SLACS and SL gene sequences at each of the eight genomic loci. We constructed overlapping subclones of the 7-kb SLACS segment from Q107 and determined whether every SLACS hybridizes to every clone. *T.gambiense* DNA was digested with *PvuII*, *PvuII* + *SalI*, *PvuII* + *BglII* and *BglII* + *Bam*HI enzymes. In Q107 these enzymes cleave in SLACS DNA as well as in the flanking SL gene sequences and generate a recognizable pattern of fragments. Genomic digests were analyzed by Southern blotting using the overlapping subclones as hybridization probes (Figure 4B).

The 4-kb *PvuII* fragment located 5′ to the last SL gene was divided into four subclones: pSB1, pBB1400, pBB600 and pSS1000 (Figure 4B). If all SL gene-carrying fragments are associated with a complete copy of SLACS, then a 4-kb *PvuII* and a 2.5-kb *PvuII* + *SalI* fragment are expected to hybridize to the most 3′ end probe, pSB1. This result is found as shown in Figure 4A, lanes 1 and 2. The adjacent 1400-bp *Bam*HI fragment also hybridizes to the same 4-kb *PvuII* fragment (Figure 4B, lane 1). In addition, this probe, pBB1400, hybridizes to a 12-kb *PvuII* fragment (marked by * on the autoradiogram) which is not found in the pSB1 hybridization. This is not due to heterogeneity in one of the SL flanking fragments, but rather to the presence of another locus (discussed below). All eight copies of SLACS, therefore, appear to contain the same size 4-kb *PvuII* fragment.

Probe pPP1600, which carries the 1600-bp *PvuII* band, 5′ to the 4-kb *PvuII* fragment discussed above, recognizes only one 1.6-kb band in genomic *PvuII* digest (Figure 4C, lane 1). The restriction enzyme maps of both phage clones show heterogeneity 5′ upstream of this *PvuII* site. Therefore, as expected, in a *Bam*HI + *BglII* digestion, five bands hybridize to probe pPP1600 (Figure 4C, lane D). In order to localize the site of heterogeneity more precisely, we have generated shorter subclones of the 4.5-kb *Bam*HI−*BglII* fragment by deleting various segments from its *Bam*HI end. One of these deletion clones, pdelBg750, contains 750 bp flanking the *BglII* site. When it is used as the probe, multiple bands (five) appear in a *PvuII* digest, which range in size from 1.8 to 3.4 kb (Figure 4D, lane 1). Based on the hybridization intensity, the 2.2-kb band represents multiple copies of this sequence. All of these *PvuII* fragments contain the *BglII* site which is also found in phage Q107 (Figure 4D, lane 3). Thus

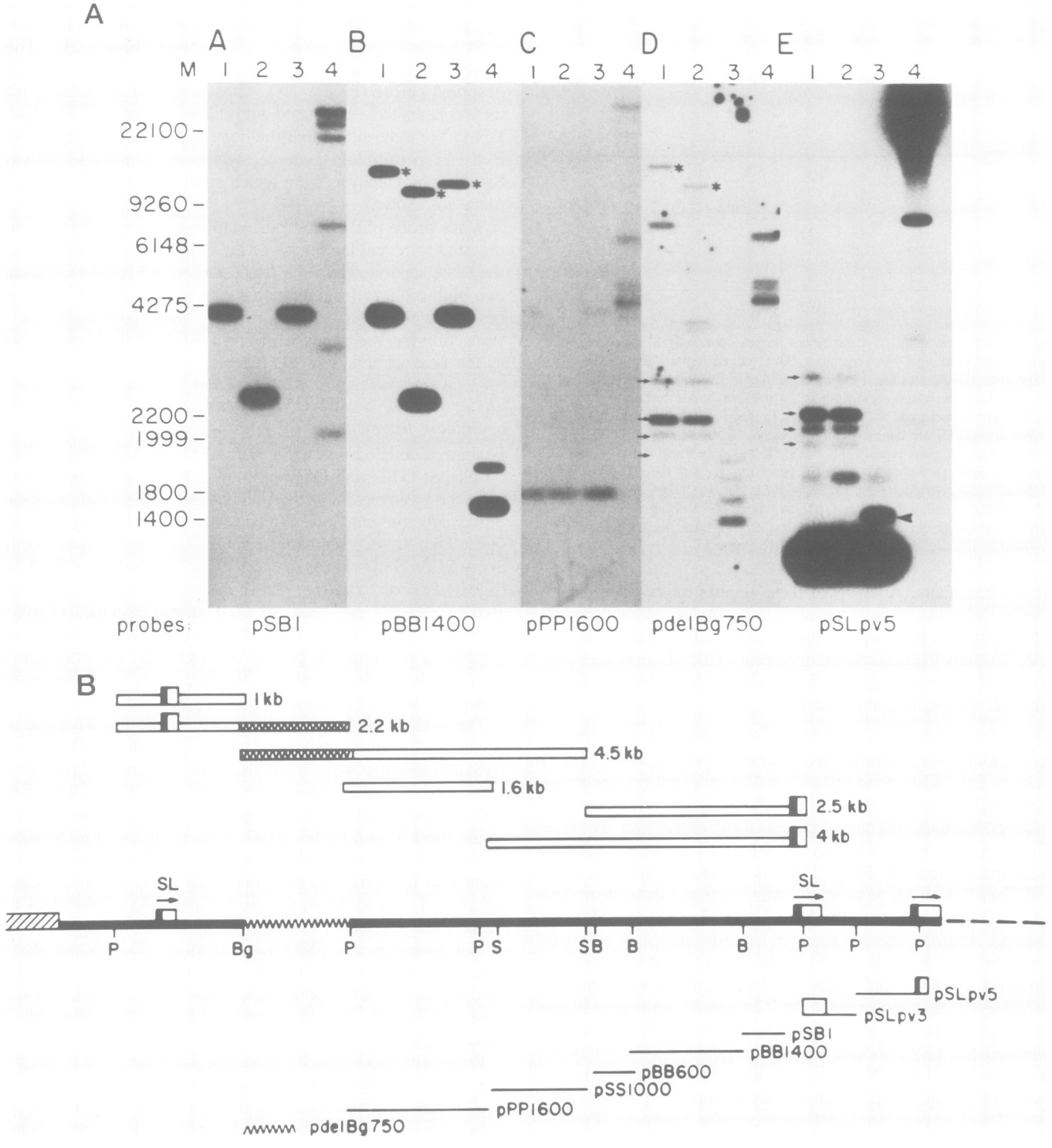

Fig. 4. Autoradiograms showing the genomic arrangement of the multiple SLACS DNA insertions. *T.gambiense* genomic DNA (2 μg) digested with *Pvu*II **(lane 1)**, *Pvu*II + *Sal*I **(lane 2)**, *Pvu*II + *Bgl*II **(lane 3)** and *Bam*HI + *Bgl*II **(lane 4)** enzymes were analyzed by gel electrophoresis on 0.6% agarose gels in TAE buffer. M denotes the lambda *Hin*dIII-cleaved DNA co-electrophoresed as the size standard. Southern blots were prepared as described in Figure 3. The probes used in each analysis are indicated below the autoradiograms. **(B)** A map of the recombinant phage Q107 from which the hybridization probes were prepared; the 4-kb *Pvu*II, 2.5-kb *Pvu*II−*Sal*I, 1.6-kb *Pvu*II, 4.5-kb *Pvu*II−*Bgl*II, 2.2-kb *Pvu*II and the 1-kb *Pvu*II−*Bgl*II fragments discussed in the autoradiograms are indicated above the recombinant phage map. In **A**, **panels B** and **D**, * marks the cross-hybridizing bands that are not directly linked to the SL gene loci. The arrows in **panels D** and **E** mark the 1800-, 2200- and 3000-bp *Pvu*II fragments that hybridize to probes pdelBg750 and to pSLpv5 respectively. In **panel E**, the arrowhead in **lane 3** denotes the 1000-bp *Pvu*II−*Bgl*II fragment that the SL gene probe recognizes in all loci. The origin of the 1.4-kb band in **panel E** is not known but it observed with all SL probes. When an SL gene 35-mer-specific nucleotide probe is used in the hybridization analysis instead of clone pSLpv5, the 4-kb *Pvu*II fragment is also observed to contain SL homologous sequences (data not shown).

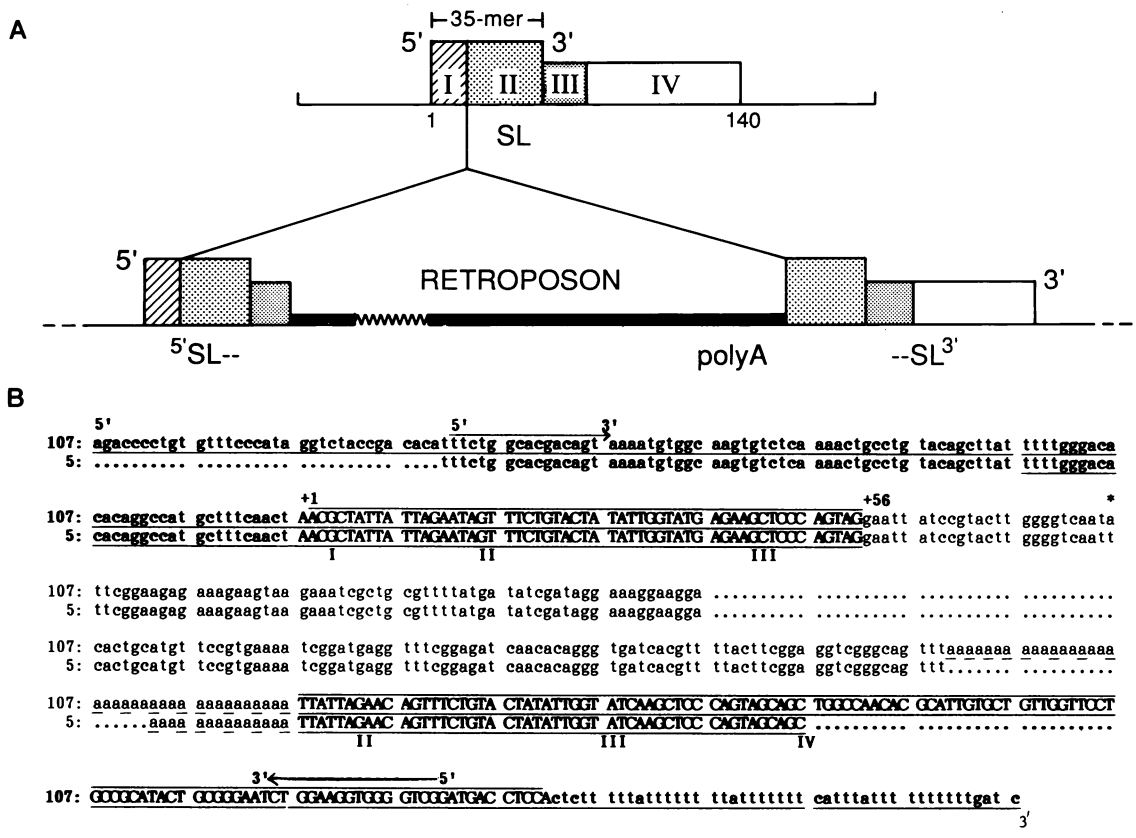

Fig. 5. The organization of one insertion element present in the SL gene cluster. (**A**) A schematic drawing showing a retroposon inserted into an SL gene repeat unit. The 140-nucleotide coding segment of the SL gene is shown by boxes. The 5′ end 35-mer sequence is comprised of boxes I and II. Box I contains the first seven nucleotides at the 5′ end, box II has the rest of the 27 bases. Box III extends from nucleotides 35 to 56 and box IV from 56 to the end of the gene, nucleotide 140. The retroposon sequence is denoted by the thicker line and interrupts the SL unit at position 7. Boxes II and III are found at both ends of the insert as direct repeats. Preceding the site of insertion at the 3′ end, the stretch of poly(A)s are shown. (**B**) The DNA sequence data corresponding to the 5′ and 3′ end borders of the SLACS insertion from both recombinant phages is shown. The SL gene repeat unit sequence is in bold lower-case letters while the 140-nucleotide SL RNA coding segment is shown in bold capital letters. The lower-case letters denote the retroposon DNA sequence. The dots represent the rest of the 7-kb retroposon sequence. 1+ marks the 5′ end of the SL RNA coding region. The 5′ end border sequence was obtained from the 1-kb *Eco*RI−*Bgl*II restriction enzyme fragment subcloned into the M13mp8 vector from both SLACS copies. The 16-mer oligonucleotide complementary to position −85 to −74 of the SL gene was used as the sequencing primer according to the dideoxy sequencing protocol (Materials and methods). The * indicates the nucleotide substitution observed between two sequences. The 3′ end border sequence from Q107 was obtained by using the synthetic oligonucleotide primer corresponding to position +105 to +121 of the SL gene. The 800-bp *Bam*HI−*Apa*I fragment spanning the retroposon and SL gene sequences was used as the template to initiate synthesis according to the dideoxy protocol using the above primer. In the case of phage Q5, the 345-bp *Bam*HI−*Pvu*II restriction enzyme fragment spanning the 3′ end SL−non SL junction (Figure 2A) was subcloned into the M13 cloning vectors and its DNA sequence information was obtained in both directions.

the variable region is present within the 1.2-kb *Pvu*II + *Bgl*II fragment of phage Q107 (Figure 4). The SLACS DNA present in phage Q5 also contains a *Pvu*II−*Bgl*II fragment in this same region, though here it is only 1 kb in length. Primary sequence analysis of clone pdelBg750 shows a 180-bp 3-fold tandem repeat (data not shown).

The 2.2-kb *Pvu*II fragment in Q107 and the corresponding 2-kb fragment in Q5 show homology with both SL and SLACS probes (data not shown). In genomic *Pvu*II digests, multiple fragments (four), hybridize to the 5′ end SL gene probe (pSLpv5) (Figure 4E, lane 1) and these correspond to the fragments homologous to probe pdelBg750 as discussed above (Figure 4D, lane 1). In the *Pvu*II−*Bgl*II digest, however, a single 1-kb fragment hybridizes to the pSLpv5 probe (Figure 4E, lane 3), thus suggesting that the 5′ end SL−non-SL junction is conserved in all of the SLACS segments.

The data therefore show that both the 3′ and the 5 ends of the eight SLACS segments border on spliced leader gene sequences. All eight SLACS sequences appear to have a common organization and each SL−non-SL junction is conserved.

In the *Pvu*II digestion analysis, probe pdelBg750 detects two additional bands of 8.0 and 8.2 kb (Figure 4D, lane 1). In PFG analysis, however, this probe hybridizes only to the two high mol. wt fragments shown in Figure 1. These 8- and 8.2-kb *Pvu*II fragments must correspond to the ends of the two 350- and 450-kb chromosomal *Hin*dIII fragments since they can be cut by *Hin*dIII (data not shown). We have isolated a recombinant phage that contains one of these regions. The restriction enzyme map of SLACS DNA in this recombinant is conserved up to ~1 kb downstream of the *Bgl*II site but diverges thereafter.

The last restriction fragment not accounted for yet, the 12-kb *Pvu*II band that also hybridizes with probe pdelBg750 (Figure 4D, lane 1), represents a locus not directly associated with the SL gene repeat units. Probes pBB1400 and pBB600 also have homology with this fragment while probes pSS1000 and pPP1600 show no hybridization (Figure 4B, lane 1, and data not shown).

This additional homology with probe pBB1400 can be located on a 2-kb *Bam*HI−*Bgl*II band (Figure 4B, lane 4) and the restriction map extending ~10 kb in each direction from this 2-kb fragment does not resemble either the SL gene repeat unit pattern or the SLACS DNA (data not shown). In PFG analysis using pBB1400 probe, we detect a 40-kb *Hin*dIII fragment in addition to the two high mol. wt fragments shown in Figure 1. However, in the *Xba*I digest, only the two high mol. wt fragments are observed and no hybridization either with the 40-kb *Hin*dIII fragment or up in the well is present (data not shown). This result suggests that the second pBB1400 and pdelBg750 homologous locus is in close proximity but not directly linked with the SL gene cluster(s).

The 7-kb SLACS is flanked by a 49 bp duplication of target DNA sequence

To understand the origin and evolution of the interrupting sequence (SLACS) that is found in the SL gene arrray, we determined the nucleotide sequence of the SL and SLACS junction segments present on phages Q107 and Q5. The SL−SLACS organization is schematically shown in Figure 5A and the DNA sequence information obtained from the junction segments is presented in Figure 5B.

At the 5′ end junction, a partial SL gene is found. It contains all of the 35-mer sequence but is followed by only 20 bases of the 140-nucleotide SL gene-coding region up to position 56 (Figure 5B). The 5′ upstream DNA sequence of this partial SL gene is identical with the SL gene repeat unit sequence. The sequence 3′ to nucleotide 56 of the SL gene, however, does not show any homology to either the 140-base SL gene-coding region or to the 1.4-kb repeat unit sequence (Figure 5B). The DNA sequence spanning the 5′ end SL−SLACS junction present on phage Q5 is found to be identical to Q107 with only one base substitution. Within the 225-bp DNA segment sequenced from this second SLACS copy, only an A was found changed to a T at position 25 in the SLACS segment 3′ to the partial SL gene (Figure 5B). Thus, at the 5′ end junction, SLACS segments interrupt the 140-nucleotide SL gene at base 56 in both of its genomic copies.

The 3′ end junction sequence shows that both SLACS copies again interrupt the SL gene-coding segment. Seven bases from the 5′ end of the 35-mer sequence are missing in both sequences (Figure 5B). The 3′ end downstream sequence, coding for the 140-nucleotide SL RNA transcript, is identical to that reported for other *T.brucei* genes. At the border of the 3′ end junction, the partial gene is preceded by a stretch of 36 As in the case of Q107 while 14 As are found in Q5. The sequence upstream of the poly(A) completely diverges from that of the SL gene repeat unit segment. This sequence is identical in both SLACS segments over the 356-bp stretch for which DNA sequencing information is obtained.

In summary, as shown in Figure 5A, the 7-kb SLACS DNA interrupts SL gene-coding segment at base 7. The 49 nucleotides between position 7 and 56 appear to have been duplicated and are present at both ends of the insert in both of its genomic copies sequenced. A stretch of poly(A) precedes the site of insertion at the 3′ end. These features are suggestive of the retroposons that have been found in eukaryotic cells.

Discussion

All spliced leader genes of *T.gambiense* are clustered on two separate loci which are present on a 350- and a 450-kb *Hin*dIII restriction enzyme fragment respectively. Multiple copies of the 7 kb (SLACS) DNA sequence interrupt the SL arrays in both loci. Due to these SLACS insertions in the tandem array of SL genes, new restriction enzyme sites are introduced which account for the previously identified SL orphons. In total, eight SLACS sequences are found associated with the SL genes in *T. gambiense*.

The primary sequence analysis of SL−SLACS junctions shows that 49 bases of the target sequence have been duplicated at the site of insertion and flank both ends of SLACS sequences as direct repeats. There is also a long consecutive run of poly(A) at one end of the insert. The 3-kb DNA region preceding the poly(A) tail contains long stretches of open reading frames (ORFs) (data not shown). These features are suggestive of the properties of retroposons found in other organisms. The hallmarks of retroposons include a poly(A) tract at the 3′ end region, flanking direct repeats associated with the transposed segment and, in higher eukaryotes, absence of introns. These features suggest that such sequences are derived from processed, spliced mRNA (Vanin, 1984, 1985). The flanking repeats are believed to be due to the mechanism of insertion. For most of the pseudogenes studied, a direct repeat of 9−14 bp has been reported. SLACS repeats are longer (49 nucleotides). However, such extensive duplications of 41 and 62 nucleotides are present in the pseudogenes of the rat (Soares *et al.*, 1985) and the human genome (Rogan *et al.*, 1987). Similar possible retroposons have been reported in *T.brucei*; RIME (Hasan *et al.*, 1984) INGI (Kimmel *et al.*, 1987) and TRS (Murphy *et al.*, 1987). The RIME mobile element sequence interrupts a ribosomal gene and insertion of this element has resulted in a 7-bp duplication of target sequences present at both ends. It has an ORF encoding a protein of 160 amino acids and the 3′ end of the element contains a stretch of 14 A residues. The other two retroposons, INGI and TRS, represent longer dispersed repetitive elements associated with RIME sequences at both ends. Both have a poly(A) tail at one end and a possible insertion site duplication. While the insertion element we report shares some of the mobile element features observed for RIME, INGE and TRS—i.e. poly(A) tail, duplication of target DNA sequence at the site of insertion—it is not a highly repetitive sequence. Unlike RIME, INGI and TRS, all of which are present in ~200−400 copies per genome, SLACS sequences are of low copy number and are only present within the SL array. Our Southern blot analysis using various probes at lower stringencies shows only one additional SLACS-related copy in the genome. This locus may represent the 'parent' sequence from which SLACS may have originated as a pseudogene. However, the overall organization of SLACS sequences is not identical with the genomic map of this SLACS-related locus. In *T.brucei*. Northern blot analysis reveals two RNA bands in the poly(A)$^+$ fraction when SLACS segments pBB1400 and pBB600 are used as probes (unpublished results). Both of these probes recognize SLACS as well as SLACS-related DNA sequences. It remains to be shown whether the transcripts we observe originate from within the SLACS segments in the SL clusters or whether they represent the SLACS-related locus. It is possible that SLACS represents an internally scrambled element as found in the *Kpn*I element of mammalian genomes (Potter, 1984) or it may contain an insertion within the main element (Rogan *et al.*, 1987). A comparison of the SLACS DNA sequence with the sequence of the related locus will distinguish between these possibilities.

Multiple SLACS segments could either represent one retroposition event which has been subsequently duplicated, or independent insertions. The DNA sequencing analysis of the SLACS−SL

junctions of two retroposons present on recombinant phages Q107 and Q5 show that both SLACS sequences interrupt the SL gene at the same site, i.e. nucleotide 7 within the 35-mer leader. Furthermore, an identical 49-bp target DNA has been duplicated and flanks both sites of insertion. This is strong evidence that the multiple SLACS copies present in the SL gene clusters originated from a single insertion. Independent multiple retroposition events would most likely have generated varying extents of target DNA duplication at different insertion sites. The few base changes observed between the two copies of the SLACS DNA sequences are compatible with genetic drift often found in a non-transcribed region. The simplest model that can account for the evolution of the spliced leader genes and the SLACS sequences is the random insertion of one sequence by retroposition into the tandem SL gene array with subsequent amplification. This could have involved unequal sister chromatid exchange (Smith, 1976) or gene conversion (Meselson and Radding, 1975).

The overall anatomy of the SL gene cluster(s) and SLACS sequences is conserved in other related trypanosome species. For instance, in the *T. brucei* serotype expressing glycoprotein (VSG) 117, we find 17 SLACS-related DNA fragments, all of which are present on an ~600-kb *Xba*I fragment (unpublished observations). Multiple SLACS sequences inserted into the SL gene cluster are also present in *T.rhodesiense* on an ~600-kb *Xba*I fragment. The flanking DNA sequence of one of the SL-containing fragments isolated from the IsTaR serodeme of *T. brucei* is in agreement with our results (Parsons *et al.*, 1986). This non-SL sequence is also found to be associated with the other SL fragments in the IsTaR serodeme, suggesting the conservation of the retroposon. One interpretation of these findings is that the retroposon insertion is an event which precedes the subspeciation of the *T.brucei* complex.

It is interesting to note that in all of three trypanosome species we have analyzed, the overall organization of SLACS is conserved. Whether SLACS is necessary either for generation of the spliced leader array or possibly for its expression, requires further study. The primary sequence analysis of the retroposon is in progress and may provide insight into its possible function.

Materials and methods

Trypanosomes

T.gambiense cloned variant antigen types of the Texas trypanozoon antigen type (Textat) serodeme were used (Merritt *et al.*, 1983). Bloodstream trypanosomes were grown in irradiated rats and cells were purified by chromatography on DEAE−cellulose (Lanham, 1968).

DNA isolation

Trypanosomal DNA was extracted by standard procedures (Bernards *et al.*, 1981). Phage recombinant DNA was isolated by the procedure of Yamamoto *et al.* (1970), and for small-scale DNA extractions the alkaline lysis procedure was used (Birnboim and Doly, 1979). For the PFG analysis trypanosomes were incorporated in 0.5% low-melting agarose blocks in PSG buffer (75 mM sodium phosphate, pH 8.0, 65 mM NaCl, 1% glucose) at a concentration of 4×10^9 cells/ml as described previously (Van der Ploeg *et al.*, 1984a; Bernards *et al.*, 1986). Blocks of agarose were transferred to lysis buffer containing 0.5 M EDTA, pH 9.5, 1% sodium-*N*-lauroylsarcosinate and 0.25 mg proteinase K per ml. Deproteinization was performed at 50°C for 48 h. Prior to digestion, blocks were washed in 200 vols of TE buffer (10 mM Tris−HCl, pH 7.5, 10 mM EDTA) containing 0.1 mM of the protease inhibitor phenylmethylsulfonyl fluoride (PMSF) for several hours at room temperature. Subsequently, blocks were equilibrated in the appropriate restriction enzyme buffer and enzyme digestions were carried out using 20−50 U of the enzyme for 6−18 h. The blocks were then inserted into the slots of a 1% agarose gel in TBE buffer (80 mM Tris−HCl, 90 mM boric acid, 2.5 mM EDTA, pH 8.3). The design of the PFG unit has been described previously (Schwartz *et al.*, 1983; Schwartz and Cantor, 1984).

Southern blots

Restriction endonuclease digestion, electrophoresis and transfer of DNA to nitrocellulose filters were performed as described previously (Bernards *et al.*, 1981). In the case of the PFG gels, the gels were soaked in 0.25 M Tris−HCl for 15 min prior to denaturation. The filters were pre-hybridized for 2−6 h.

Construction and screening of genomic clones

The Textat I DNA library was a gift from Christian Tschudi (Tschudi *et al.*, 1985). Briefly, high mol. wt DNA from *T.gambiense* variant I was partially digested with the restriction enzyme *Sau*3A. Fragments ranging in size from 18 to 24 kb were isolated on a 5−24% NaCl gradient and were cloned into *Bam*HI-cut bacteriophage vector EMBL4. Ligation products were packaged *in vitro* and plated on a lawn of Q358 cells. For the isolation of recombinant phages Q107 and Q5, filters were prepared from the DNA library (Benton and Davis, 1977) and were screened with the SL gene probe pSLpv5 (construction described below). [^{32}P]dATP-labelled transcripts were synthesized by Klenow enzyme *in vitro* using single-stranded template DNA from pSLpv5 according to a modified DNA-sequencing protocol (Hu and Messing, 1982). Following annealing of 10 ng M13 Universal primer to 0.5 μg template DNA at 65°C for 10 min, 37°C for 20 min and at room temperature for 10 min in primer buffer (5 ×; 0.25 M Tris−HCl pH 8.1, 10 mM dithiothreitol, 25 mM $MgCl_2$, 0.1 M KCl), unlabelled dNTPs were added (40 μM final concentration) along with 50 pmol of [^{32}P]dATP and 5 U of Klenow enzyme. Reaction was carried at 37°C for 60 min. Unincorporated labelled nucleotides were separated by ethanol precipitation and 1×10^6 c.p.m. of the labelled probe was used per ml of hybridization solution. Hybridization conditions were 50% formamide, 5 × Denhardt's (0.1% Ficoll, 0.1% polyvinylpyrrolidone, 0.1% BSA), 5 × SSC (1 × SSC is 180 mM NaCl, 10 mM NaH_2PO_4 pH 7.4,1 mM EDTA) and 0.1% SDS at 42°C. The final stringency condition of the post-hybridizatiuon wash was 0.2 × SSC, and 0.1% SDS at 65°C.

Construction of hybridization probes

To construct the SL gene probe, pSLpv5, Textat I DNA was cleaved with *Pvu*II. Restriction fragments ranging from 0.5 to 1 kb were size fractionated on a 1.2% agarose gel and were purified by the gel-electroelution approach (McDonnell *et al.*, 1977). DNA was ligated into the M13mp8 cloning vector and 71.18 cells were transformed (Messing, 1983). Recombinant plaques were screened with [γ-^{32}P]ATP (NEN) labelled (Maniatis *et al.*, 1982) SL gene-specific nucleotide probe, 5′-AACGCTATTATTAGAAC-3′, synthesized by the procedure of Van der Marel *et al.* (1982). Single-stranded template DNA was prepared from positive plaques (Sanger *et al.*, 1980) and recombinants were further characterized by DNA sequencing following the dideoxy chain termination method with the modification that [^{35}S]dATP was used for the reactions (Sanger *et al.*, 1977).

In order to construct probes pSB1, pBB1400, pBB600, pSS1000, pPP1600, the corresponding restriction enzyme fragments were purified from the phage Q107 by electroelution following gel electrophoresis (McDonnell *et al.*, 1979). These fragments were subcloned into the M13mp8 and M13mp9 vectors cleaved with the appropriate enzymes.

Probe pdelBg750 was generated using the deletion subcloning approach (Dale *et al.*, 1985). The 4.5-kb *Bam*HI−*Bgl*II fragment from Q107 was subcloned into *Bam*HI-cleaved M13mp8 DNA. Single-stranded DNA from this subclone was used to generate deletion clones using the cyclone system kit obtained from International Biotechnologies Inc. Briefly, the oligomer RD29-mer was annealed to 4 μg of single-stranded template DNA at the 3′ end of the cloned insert, in buffer containing 33 mM Tris−acetate, pH 7.8, 6.6 mM potassium acetate and 10 mM magnesium acetate at 60°C for 30 min. The annealed template primer was linearized with *Hin*dIII enzyme at 37°C for 2 h. The 3′ to 5′ specific exonuclease activity of T4 DNA polymerase was employed to digest the linearized DNA from the 3′ end. Aliquots were removed at intervals of 5, 10, 15, 30, 45 and 60 min, heat inactivated and pooled. The deletion products were tailed with dAs at 37°C for 10 min using 15 U of terminal deoxyribonucleotidyltransferase (Tdt). Fresh oligomer RD29-mer was annealed to the deletion products joining the two ends of the molecule. The remaining nick was sealed with T4 DNA ligase and competent 71.18 cells were used for transformation. The transformants were grown in 1 × YT medium for 10−12 h and 25 μl of the supernatant was mixed with 5 μl of SDS loading dye (10 × is 250 mM Na_3EDTA pH 8.0, 1% SDS, 0.1% bromophenol blue and 50% glycerol) and the size of the recombinant phages was analyzed by electrophoresis on a 0.7% agarose gel in TBE buffer (90 mM Tris−borate, 90 mM boric acid, 2.5 mM EDTA, pH 8.3) to determine the extent of deletion. One such clone, pdelBg750, was estimated to contain a 750-bp DNA insert and was subsequently purified and further analyzed.

DNA sequence analysis

To obtain the 5′ end border sequence, the most 5′ 1-kb *Eco*RI−*Bgl*II fragment from phages Q107 and Q5 was purified and subcloned into the *Eco*RI−*Bam*HI-cut M13mp9 vector. Single-stranded template DNA was purified from the recombinant white plaques and the border DNA sequence was obtained by the dideoxy approach (Sanger *et al.*, 1977). An oligonucleotide 5′-TTTCTGGCA-CGACAGT-3′ was synthesized that is homologous to SL gene upstream sequences (−85 to −69) and was subsequently used as the sequencing primer.

The 3′ end border sequence was determined using an M13mp8 recombinant clone which contains the 800-bp *Bam*HI−*Apa*I restriction spanning the SL−non-SL junction in phage Q107. A 2-kb *Apa*I fragment containing the 3′ end border

sequence was purified and its *Apa*I ends were filled in using dCTP and the large fragment of DNA polymerase I. The blunt-ended fragment was cleaved with *Bam*HI and was ligated into *Bam*HI–*Sma*I-cut M13mp9 vector. Single-stranded DNA was purified from the clone containing the 800-bp fragment and a 3′ end SL gene-specific nucleotide probe, 5′-CTGGGGTGGAAGGTCT-3′ (+105 to +121) was used to determine the 3′ end junction sequence by the dideoxy approach.

Acknowledgements

We are grateful to Drs Elisabetta Ullu and Christian Tschudi for helpful comments and to Dr Martine Y.K.Armstrong, Suzanne Wait and Louise C.Benson for their assistance and to Charlene Sullivan for manuscript preparation. This work was supported by The MacArthur Foundation and grants AI-08614 (FFR) and AI-21784 (LHTvdP) form the USPHS. S.A., T.M.L., L.H.T.V.d.P. and F.F.R. are Investigators of the Consortium on the Biology of Parasitic Diseases of The MacArthur Foundation.

References

Bernards,A., Van der Ploeg,L.H.T., Frasch,A.C.C., Borst,P., Boothroyd,J.C., Coleman,S. and Cross,G.A.M. (1981) *Cell*, **27**, 497–505.

Bernards,A., Kooter,J.M., Michels,P.A.M., Moberts,R.M.P. and Borst,P. (1986) *Gene*, **42**, 313–322.

Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–182.

Birnboim,H.C. and Doly,J. (1979) *Nucleic Acids Res.*, **7**, 1513–1523.

Borst,P. (1986) *Annu. Rev. Biochem.*, **55**, 701–732.

Campbell,D.A., Thornton,D.A. and Boothroyd,J.C. (1984) *Nature*, **311**, 350–355.

Dale,R.M.K., McClure,B.A. and Houckins,J.P. (1985) *Plasmid*, **13**, 31–40.

DeLange,T., Liu,A.Y.C., Van der Ploeg,L.H.T., Borst,P., Tromp,M.C. and Van Boom,J.H. (1983) *Cell*, **34**, 891–900.

Hasan,G., Turner,M.J. and Cordingley,J.S. (1984) *Cell*, **37**, 333–341.

Hu,N.-T. and Messing,J. (1982) *Gene*, **17**, 271–277.

Kimmel,B.E., Ole-Moiyoi,O.K. and Young,J.R. (1987) *Mol. Cell. Biol.*, **7**, 1465–1475.

Kooter,J., DeLange,T. and Borst,P. (1984) *EMBO J.*, **3**, 2387–2392.

Lanham,S.M. (1968) *Nature*, **218**, 1273–1274.

Maniatis,T., Frisch,E.F. and Sambrook,J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McDonnell,M.W., Simon,M.N. and Studier,F.W. (1977) *J. Mol. Biol.*, **110**, 119.

Meselson,M.S. and Radding,C.M. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 358–361.

Messing,J. (1983) *Methods Enzymol.*, **101**, 28–78.

Merritt,S.C., Tschudi,C., Konigsberg,W.H. and Richards,F.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1536–1540.

Millhausen,M., Nelson,R.G., Sather,S., Selkirk,M. and Agabian,N. (1984) *Cell*, **38**, 721–729.

Murphy,N.B., Pays,A., Tebabi,P., Coquelet,H., Guyaux,M., Steinert,M. and Pays,E. (1987) *J. Mol. Biol.*, **195**, 855–871.

Nelson,R.G., Parsons,M., Barr,P.J., Stuart,K., Selkirk,M. and Agabian,N. (1983) *Cell*, **34**, 901–909.

Parsons,M., Nelson,R.G., Stuart,K. and Agabian,N. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 684–688.

Parsons,M., Nelson,R.G. and Agabian,N. (1986) *Nucleic Acids Res.*, **14**, 1703–1718.

Potter,S. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 1012–1016.

Rogan,P.K., Pan,J. and Weissman,S.M. (1987) *Mol. Biol. Evol.*, in press.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161–178.

Schwartz,D.C. and Cantor,C.R. (1984) *Cell*, **37**, 67–75.

Schwartz,D.C., Saffran,W., Welsch,J., Hass,R., Goldenberg,M. and Cantor,C.R. (1983) *Cold Spring Harbor Symp. Quant. Biol.*, **47**, 189–195.

Shea,C., Glass,D.J., Parangi,S. and Van der Ploeg,L.H.T. (1986) *J. Biol. Chem.*, **261**, 6056–6063.

Smith,G.P. (1976) *Science*, **191**, 528–535.

Soares,M.B., Schon,E., Henderson,A., Karathanasis,S.K., Catl,R., Zeitlin, S., Chirgwin,J. and Efstratiadis,A. (1985) *Mol. Cell Biol.*, **5**, 2090–2103.

Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503–517.

Tschudi,C., Young,A.S., Ruben,L., Patton,C.L. and Richards,F.F. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 3998–4002.

Van der Marel,G.A., Marugg,J.E., DeVroom,E., Wille,G., Tromp,M., Van Boeckel,C.A.A. and Van Boom,J.H. (1982) *Rec. Trav. Chim. Pays-Bas*, **101**, 239–241.

Van der Ploeg,L.H.T. (1986) *Cell*, **47**, 479–480.

Van der Ploeg,L.H.T., Schwartz,D.C., Cantor,C.R. and Borst,P. (1984a) *Cell*, **37**, 77–84.

Van der Ploeg,L.H.T., Cornelissen,A.W.C.A., Michels,P.A. and Borst,P. (1984b) *Cell*, **39**, 213–221.

Vanin,E.F. (1984) *Biochim. Biophys. Acta*, **782**, 231–241.

Vanin,E.F. (1985) *Annu. Rev. Genet.*, **19**, 253–272.

Yamamato,K.R., Alberts,B.M., Berzinger,R., Lawhorne,L. and Treiber,G. (1970) *Virology*, **40**, 734.

Received on June 24, 1987; revised on September 14, 1987