Volume 16 Number 15 1988 Nucleic Acids Research

Discontinuous transcription in *Leptomonas seymouri*: presence of intact and interrupted mini-exon gene families

Vivian Bellofatto*, Robin Cooper and George A.M.Cross

The Rockefeller University, 1230 York Avenue, New York, NY 10021-6399, USA

Received April 27, 1988; Revised and Accepted July 6, 1988 Accession nos X07487, X07488

ABSTRACT

Mature mRNAs of trypanosomatid protozoa result from the joining of at least two exons, which are initially transcribed as separate RNAs. In all trypanosomatids examined to date, the first exon (mini-exon) is encoded by approximately 200 tandemly reiterated genes. In characterizing the mini-exon genes of *Leptomonas seymouri*, we identified two predominant size classes of repetitive sequences that hybridized strongly to the *L. seymouri* mini-exon sequence. These two sequences are arranged as interspersed clusters. DNA sequence analysis of a clone representing the smaller size class demonstrated that these sequences have the capacity to encode a mini-exon donor (med)RNA corresponding to the 86 nt component seen in Northern blots of *L. seymouri* RNA. The larger size class comprises a family of related sequences, some of which contain DNA inserted into the mini-exon portion of the medRNA gene. The specific insert identified here (LINS 1) is exclusively associated with medRNA sequences, and is present in approximately 20% of the larger size class of *L. seymouri* medRNA genes. Disregarding the insertion, the sequences of the smaller *bona fide* mini-exon genes and the gene copy containing the insert were almost identical. The insert sequence is transcribed in the same direction as medRNA to yield at least four small non-polyadenylated RNAs, which appeared not to be linked to medRNA sequences.

INTRODUCTION

The discontinuous synthesis of mRNA is a distinct feature of the trypanosomatid flagellate protozoa. A mature RNA results from the joining of at least two exons, which are initially transcribed as two separate RNAs (1). Production of mRNA in this way is not limited to these organisms since an analogous situation has been recently detected in the metazoan *Caenorhabditis elegans* (2).

In *Trypanosoma brucei*, the 5' exon common to all characterized mRNAs is initially transcribed as a short non-polyadenylated RNA, termed the mini-exon-donor RNA (medRNA), because its 5' end, which constitutes the mini-

exon, is a 39 nt sequence that is spliced onto the coding exon during mRNA maturation. The mini-exon sequence is sufficiently conserved among trypanosomatids to allow the identification of mini-exon genes in many genera by Southern blot analysis using a *T. brucei* mini-exon sequence as probe. Such experiments revealed that, in all species except *T. congolense* (3), a direct and tandemly repeated unit of a single size encodes the medRNA. These repeated units are a different size in each species and, where characterized, have been shown to be divergent in sequence except for limited regions within and immediately 3' to the medRNA transcript.

The approximately 200 copies of the medRNA gene repeat unit are tandemly arrayed at a limited number of genomic locations, although orphon copies do occur (4-7). It was shown recently, in *T. brucei* and *T. gambiense*, that specific DNA elements surround medRNA gene arrays and orphons (8,9). These elements, termed the medRNA gene associated elements (MAEs) (9), or spliced-leader-associated conserved sequences (SLACS) (8), vary between 5.5 and 7 kb in length and were only found associated with medRNA genes. In one case, a SLACS resembled a retroposon because it contained a terminal poly (A) tract and a 49 bp duplication of target DNA at the integration site.

We have found that the mini-exon genes of *L. seymouri* are contained within three related gene families. The family represented by 0.75 kb *Hin*dIII fragments most likely encodes an 86 nt medRNA, which contains the 5' donor exon used in the maturation of *L. seymouri* mRNAs. *Hin*dIII fragments 1.1 kb in size define a second family, members of which contain a localized region of non-homology to the 0.75 kb *Hin*dIII family, which accounts for the difference in repeat-unit length between the two families. In the clone we have analyzed, the non-homologous region is a conserved 293 bp sequence present in approximately 20% of the 1.1 kb *Hin*dIII family members. The insert is represented in at least four non-polyadenylated transcripts. A third family, composed of 5.9 kb *Hin*dIII fragments, contains mini-exon and insertion sequences in unknown arrangements.

MATERIALS AND METHODS

Leptomonas seymouri (ATCC 30220), originally isolated from the hemipteran *Dysdercus suturellus* and subsequently cloned by us, was maintained in logarithmic growth, between 10^5 and 3×10^7 cells/ml, at 27°C in Bone and Steinert's medium (10).

DNA was isolated as described (11). DNA probes were labeled with ^{32}P α-dCTP using the random hexamer priming method of Feinberg and Vogelstein

(12). Oligonucleotides were labeled with ^{32}P γ-ATP and T4 polynucleotide kinase (13). The *L. seymouri* mini-exon sequence was determined by primer extension of the 5' end of α-tubulin mRNA (14) . Clones pM4 and pM8, containing the mini-exon 0.75 and 1.1 kb repeat units, were obtained by screening size-fractionated *L. seymouri Hin*dIII restriction fragments, inserted in pGEM-3 (Promega Biotec), with a ^{32}P-labeled 35 nt synthetic oligonucleotide (MX oligonucleotide) that complements the mini-exon sequence of *L. seymouri* (Figure 2A). All plasmids were amplified in *E. coli* DH1 cells and colonies were screened according to Grunstein and Hogness (15). Filters containing colonies were hybridized at 37°C in 5x SSC, 5x Denhardt's solution, 50 mM Na phosphate buffer pH 6.8, 50% formamide, and 250 µg/ml denatured, sheared salmon sperm (ss) DNA and washed in lx SSC, 0.1% SDS at 25°C (13). Subcloning into pGEM-3 was done using standard procedures. pLINS 1 was generated by cloning the 225 bp *Hin*cII-*Fsp*I fragment from pM8 into the *Sma*I site of pGEM3. Sequencing was done using the chain termination method of Sanger *et al.* (16), with the modifications described by Tabor and Richardson (17). Pairwise sequence alignments were compiled on a VAX 11/780 computer using the ARPMON software package (18).

For Southern blots, DNA, size-fractionated by electrophoresis on agarose gels, was transferred onto Nytran (Schleicher and Schuell) and hybridized in 50% formamide, 1x Denhardt's solution, 1% SDS, 100 µg/ml ssDNA and 6x SSPE (13) at 42°C. Washes were in 0.2x SSC, 0.1% SDS at 55°C. DNA for slot blots was denatured (14) and hybridizations were done at 37°C in 50% formamide, 2x Denhardt's solution, 0.1% SDS, 100 µg/ml ssDNA, 5x SSC and washed in 0.2x SSC, 0.1% SDS at 37°C.

RNA was prepared from cultures at 5 x 10^6 cells/ml either by guanidinium isothiocyanate-SDS lysis or 65°C phenol-SDS extraction (13). Poly(A)$^+$ RNA was isolated by two passages through an oligo dT-cellulose column (11). RNA-denaturing polyacrylamide gels were electroblotted onto Nytran using procedures provided by the manufacturer. Blots were hybridized at 50°C in 50% formamide, 5x SSPE, 200 µg/ml ssDNA, 2 x Denhardt's solution, 0.5% SDS and washed in 0.1x SSC, 0.1% SDS at 60°C before autoradiography.

^{32}P α-UTP-labeled riboprobes were synthesized from linearized pGEM-3-derived clones as described by Promega Biotec protocols.

RESULTS

Cloning of mini-exon genes

The sequence of the primer extension product derived from the 5' end of

α-tubulin mRNA revealed a 35 nt 5' terminal RNA sequence that was not present within the α-tubulin genomic DNA repeat unit (14). Similarity to the 5' end of other trypanosomatid mRNAs was sufficient to allow tentative designation of this 35 nt sequence as part of a probable 39 nt (19,20) *L. seymouri* mini-exon sequence . To clone DNA encoding the *L. seymouri* mini-exon sequence, we initially hybridized a Southern blot of restriction enzyme-digested genomic DNA with a ^{32}P-labeled 35 nt oligonucleotide complementary to the *L. seymouri* mini-exon RNA sequence. Two prominent bands of 1.1 and 0.75 kb, and a faint 5.9 kb band, were observed by hybridization to a *Hin*dIII digest of genomic DNA (Figure 1A). Bands of 1.1 and 0.75 kb were also obtained in *Sph*I, *Ava*I, *Bgl*I, *Hin*fI, *Rsa*I , and *Mae*III digests (Figure 1B). Therefore, there appeared to be two major classes of reiterated mini-exon genes. Unit-length fragments from each class delineated by *Hin*dIII were cloned in pGEM 3 for further analyses.

The 0.75 kb repeat unit encodes *L. seymouri* medRNA

pM4, a clone containing a representative copy of the 0.75 kb mini-exon-containing repeat, as deduced from restriction map comparisons between several isolated clones, was sequenced by the dideoxy chain termination method (Figure 2A). An intact mini-exon sequence is present within pM4. Figure 1B shows that enzymes which cut once within the 0.75 kb of *L. seymouri* DNA present in pM4 generate 0.75 kb mini-exon-containing fragments in genomic digests (lanes 1,2,5,6,7,12). Thus, pM4 represents an unrearranged copy of the 0.75 kb mini-exon gene. The 0.75 kb *Rsa*I fragments (lane 10) were unexpected in the total *Rsa*I digest because there are two *Rsa*I sites in pM4 (Figure 2B). It is likely that there is microheterogeneity at one *Rsa*I site in the 0.75 kb mini-exon genes in the cell, most likely not at the site within the mini-exon sequence, where a mutation might destroy mini-exon function.

A ^{32}P-labeled riboprobe, transcribed from the *L. seymouri* DNA within pM4, detected a non-polyadenylated 86 nt RNA on Northern blots (Figure 3). The MX oligonucleotide probe also hybridized to the 86 nt RNA. A riboprobe from the complementary strand of pM4 and an oligonucleotide equivalent to the mini-exon sequence did not hybridize to any RNA. Therefore, transcription within the 0.75 kb gene represented on pM4 is unidirectional and yields a single, nonpolyadenylated mini-exon-containing 86 nt RNA.

An 86 nt RNA containing the mini-exon sequence at its 5' end would terminate 15 nt prior to the stretch of T's present in the pM4 sequence (see Figure 2A). This characteristic, as well as small RNA size, lack of polyadenylation, and the presence of the mini-exon sequence, identifies this 86

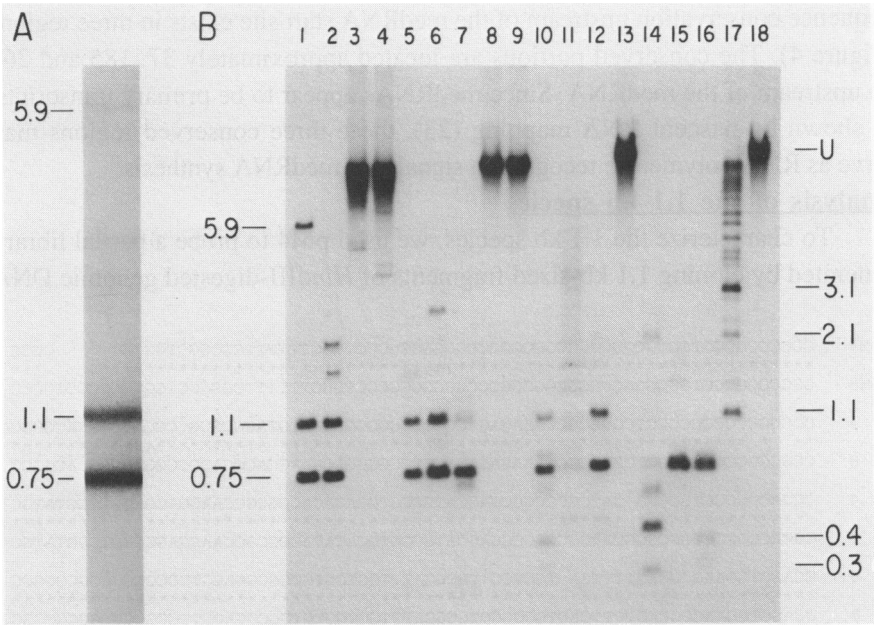


Figure 1. (A) *L. seymouri* DNA, digested with *Hin*dIII, electrophoresed through a 1.5% agarose gel, transferred to nitrocellulose and probed with the mini-exon oligonucleotide (MX oligonucleotide). (B) Southern blot of multiple digests of *L. seymouri* DNA probed with the *L. seymouri* DNA contained within pM4. Digests are: *Hin*dIII (lane 1), *Sph*I (lane 2), *Pst*I (lane 3), *Pvu*II (lane 4), *Ava*I (lane 5), *Bgl*I (lane 6), *Hin*fI (lane 7), *Bgl*II (lane 8), *Cla*I (lane 9), *Rsa*I (lane 10), *Ban*I (lane 11), *Mae*III (lane 12), *Apa*I (lane 13), *Hin*dIII/*Sph*I (lane 14), *Hinc*II/*Hin*dIII (lane15), *Hinc*II/*Rsa*I (lane 16), *Hinc*II (lane 17), and uncut DNA (lane 18). All digests are complete except the *Apa*I and *Hin*dIII/*Sph*I. Sizes are in kb. U is uncut DNA.

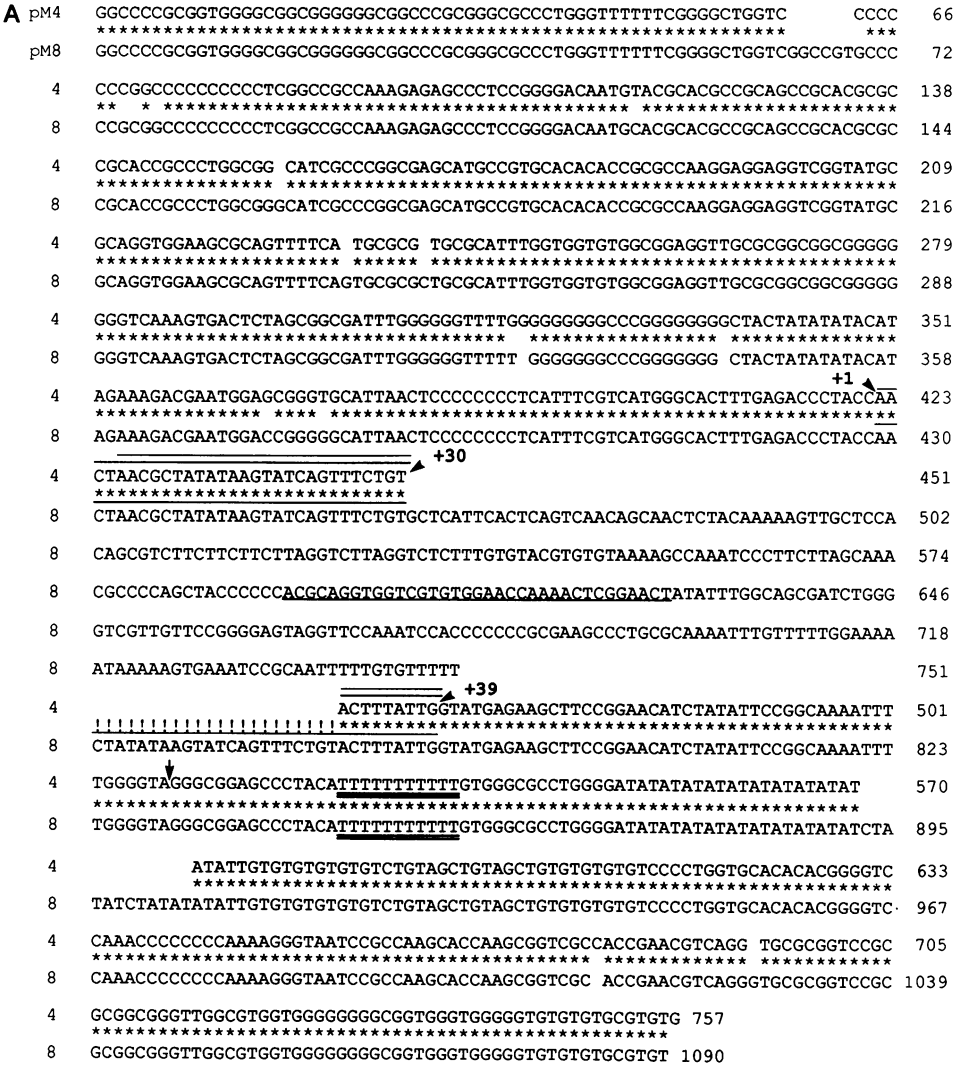
nt RNA as the *L. seymouri* medRNA. By analogy with *T. brucei*, the invariant AACT sequence preceeding the 35 nt sequence determined by primer extension of the 5' end of α-tubulin mRNA (14) is probably transcribed and extensively methylated as part of the RNA cap structure (19-21).

Sequences upstream from the medRNA coding region were analyzed for potential transcription initiation signals by comparing 300 bp of upstream sequence from *L. seymouri* with an equivalent length of sequence 5' to the medRNA transcripts in other trypanosomatids. Comparisons were limited to 300 bp because that is the length of the nontranscribed region in the smallest known mini-exon gene repeat identified thus far in a trypanosomatid (6,22).

Sequence conservation upstream of the medRNA start site exists in three regions (Figure 4). The conserved portions are located approximately 37, 185 and 265 bp upstream of the medRNA. Since medRNAs appear to be primary transcripts, as shown by nascent RNA mapping (23), these three conserved regions may serve as RNA polymerase recognition signals for medRNA synthesis.

Analysis of the 1.1 kb species

To characterize the 1.1 kb species, we used pM4 to probe a partial library generated by cloning 1.1 kb-sized fragments of *Hin*dIII-digested genomic DNA



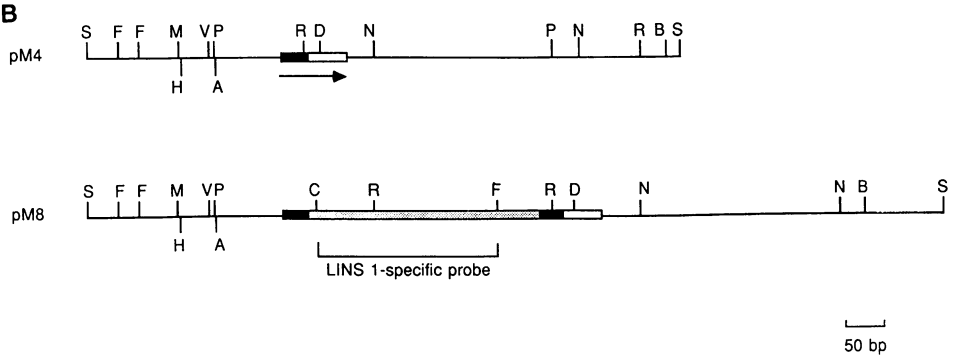


Figure 2. (A) DNA sequence of pM4 and pM8. Nucleotide identity is indicated by * . The overline shows the probable 39 nt mini-exon, which also is numbered in bold type. The "MX oligonucleotide" complements nt 4 through 39, which are labeled with a double overline. The arrow denotes the approximate 3' end of the medRNA. The directly repeated portion of the mini-exon sequence that flanks the LINS 1 insert in pM8 is indicated by ! The thick underline shows the stretch of T's, which follow the 3' end of the medRNA in all trypanosomatids. The thin underline indicates the complement of the LINS 1-specific oligonucleotide used in hybridizations. (B) The restriction maps of the *L. seymouri* DNA in pM4 and pM8. A repeat unit is shown between the *Sph*I sites. The black box represents all or part of the 39 bp mini-exon sequence and the open box shows the remaining portion of the medRNA-encoding DNA. The stippled box shows the 293 bp LINS 1 sequence, part of which was subcloned to generate pLINS1 (labeled "LINS 1-specific probe"). Transcription is from left to right for the medRNA (indicated by the arrow) and LINS 1 RNAs. Restriction sites are: *Sph*I (S), *Mae*III (M), *Ava*I (V), *Bgl*I (B), *Hin*fI (H), *Hin*cII (C), *Rsa*I (R), *Sma*I (A), *Hin*dIII (D), *Fsp*I (F), *Apa*I (P), and *Ban*I (N).

into pGEM 3. Ten independent clones were identified and partial sequence data from four clones confirmed homology to a part of the medRNA coding region found in pM4. Two of these four clones had identical restriction maps. One of these, designated pM8, was completely sequenced (Figure 2A). The sequence of pM8 is nearly identical to that of pM4 except that the mini-exon sequence in pM8 is interrupted at nucleotide 458 by a 293 bp insert that is followed by a duplication of bases 9 through 30 of the mini-exon sequence. We refer to the insert as LINS 1 because it is the first identified *L. seymouri* insertion sequence type 1. Hybridization of a LINS 1-specific probe to the ten 1.1 kb class clones detected the presence of three types of inserts: those that hybridized strongly, weakly or not at all.

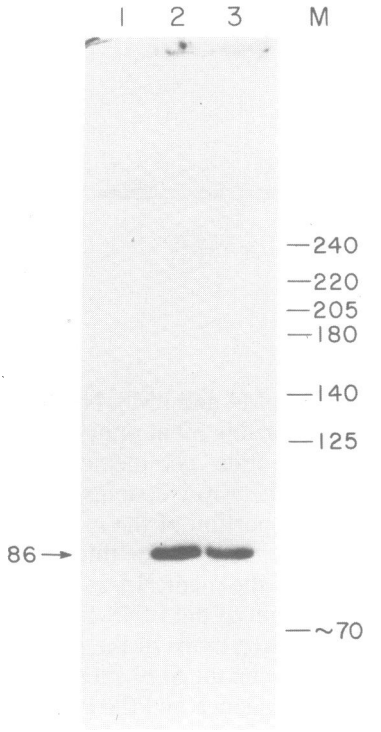
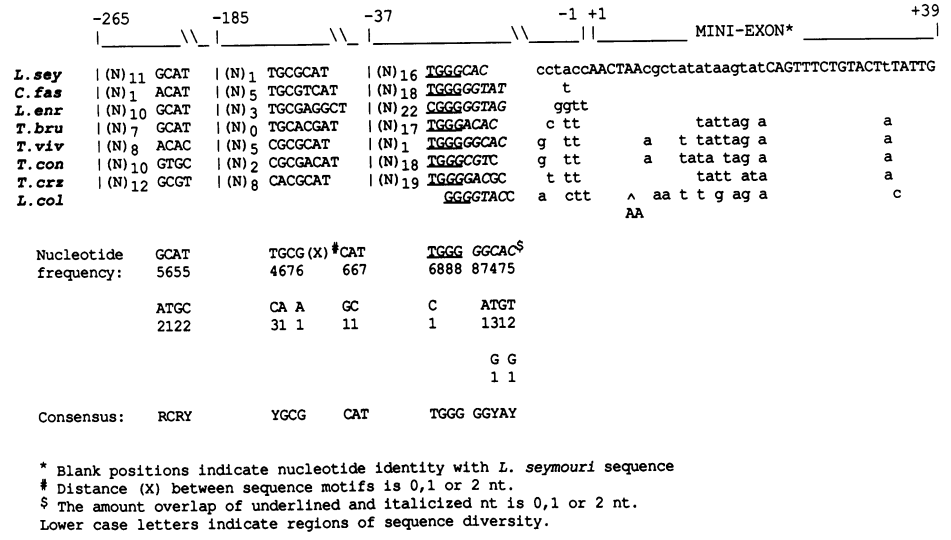


Figure 3. Northern blot analysis of medRNA. 3 μg of poly (A)$^+$ RNA, 6 μg total RNA or poly (A)$^-$ RNA (lanes 1,2,3 respectively) were electrophoresed on a 7M urea-8% polyacrylamide gel, transferred to Nytran, and hybridized with a ^{32}P-labeled riboprobe transcribed from pM4 and containing the complement of the mini-exon sequence. Markers show the sizes of the six small rRNAs in *L. seymouri*. These sizes were obtained by comparing the *L. seymouri* small rRNAs with *T. brucei* small rRNAs of known size (24,25). The 240 nt RNA marker is from the Bethesda Research Laboratory RNA marker kit. The exposure was overnight using two intensifying screens. On longer exposures, a weak band of approximately 95 nt appears in lanes 2 and 3. Under low stringency hybridization conditions (those that would specifically hybridize 26 nt of the MX oligonucleotide to RNA) and extended autoradiography, an identical pattern to that shown here was observed except that some hybridization to the 220 nt rRNA was detected. Sizes are in nucleotides.

The presence of the *Rsa*I 0.24 kb band (clearly visible on the original autoradiogram) and the *Hin*dIII/*Hin*cII 0.32 kb band (Figure 5, lanes 10 and 15), predicted from the sequence of pM8 and the restriction map of an independently



* Blank positions indicate nucleotide identity with *L. seymouri* sequence
\# Distance (X) between sequence motifs is 0,1 or 2 nt.
$ The amount overlap of underlined and italicized nt is 0,1 or 2 nt.
Lower case letters indicate regions of sequence diversity.

Figure 4. Sequences upstream from and containing the mini-exon sequence are compared among eight different trypanosomatids. The consensus sequences contain bases that are present in six out of seven, or seven out of eight times at a given position. Sequence data are from this work and references 3,6,7,22,23,26-28. ***L.sey*** (*L. seymouri*); ***C.fas*** (*C. fasciculata*); ***L.enr*** (*Leishmania enriettii*); ***T.bru*** (*T. brucei*); ***T.viv*** (*T. vivax*); ***T.con*** (*T. congolense*); ***T.crz*** (*T. cruzi*); ***L.col*** (*Leptomonas collosoma*).

isolated clone identical to pM8, confirmed that the genomic arrangement of LINS 1 is accurately represented in pM8. To characterize this insert we synthesized a 35 nt oligonucleotide (LINS 1-oligonucleotide) equivalent to the central portion of LINS 1 and different from any sequence in pM4 (the oligonucleotide complements the underlined sequence in Figure 2A). Figure 6, bottom panel, shows the result of hybridization of the LINS 1-oligonucleotide probe to slot blots of genomic and pM8 DNA. These data indicated that there are approximately 60 copies of LINS 1 per cell. Hybridization using a ^{32}P-labeled DNA from the *L. seymouri* sequence within pLINS 1 (see Figure 2B) to probe *Hin*dIII-digested genomic DNA showed that copies of the LINS 1 sequence are distributed equally between and exclusively to the 1.1 and 5.9 kb species (Figure 7C and Figure 5). Thus, there are approximately 30 copies per cell of 1.1 kb *Hin*dIII fragments that contain LINS 1. As detailed below, there are approximately 140 copies per cell of 1.1 kb *Hin*dIII fragments that contain mini-exon sequences. Hybridization of LINS 1-specific and mini-exon-specific probes

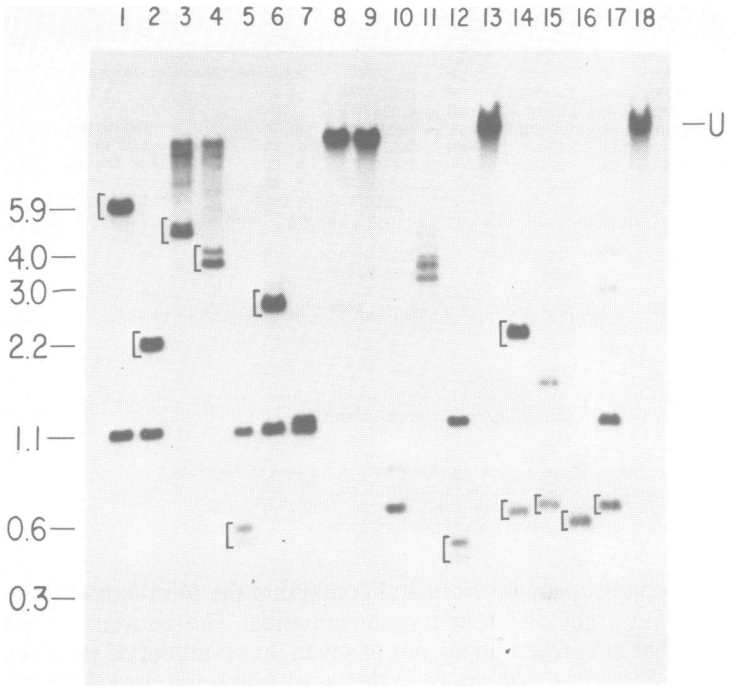


Figure 5. Southern blot showing genomic organization of LINS 1-hybridized sequences. The blot shown in Figure 1B was washed to remove the probe used previously and rehybridized with the *L. seymouri* DNA insert of pLINS 1. Based on hybridization intensity, the bracketed bands are most likely fragments that overlap the 5.9 kb *Hin*dIII species. Sizes are in kb.

to genomic 1.1 kb fragments produced by several different restriction enzyme digests demonstrated that the two sequences were adjacent to each other in the genome (Figure 1B and 5), as in pM8. These copy number and sequence distribution analyses indicated that approximately 20% of the 1.1 kb mini-exon - containing species contain LINS 1. Strong hybridization by LINS 1 to two out of the ten mini-exon-containing 1.1 kb *Hin*dIII fragment clones was consistent with this.

Quantitation of mini-exon containing-genes in *L.seymouri*

Genomic and pM4 DNAs were hybridized with the MX oligonucleotide to determine the copy number of mini-exon-containing genes in *L. seymouri*. Hybridizations showed that there are approximately 425 copies of the mini-exon sequence per cell (Figure 6). To determine the ratio of 0.75, 1.1 and 5.9 kb mini-exon-containing species to each other, and thus ascertain the approximate

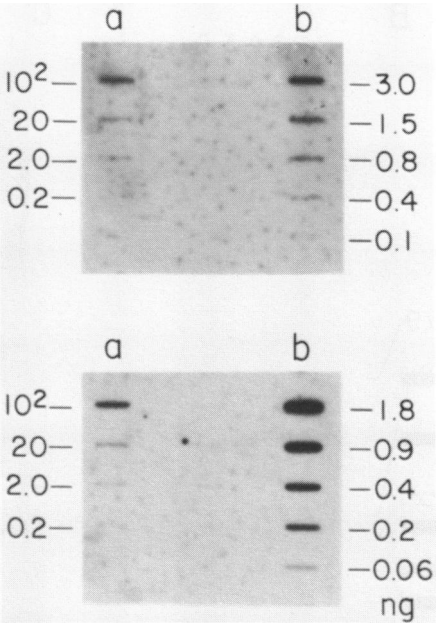


Figure 6. Slot blots to determine mini-exon and LINS 1 copy number in the genome. Top panel: (a) DNA from given number of *L. seymouri* cells x 10^{-4} (b) amount of pM4 DNA (ng). Probe was radiolabeled MX oligonucleotide. Bottom panel: (a) DNA from given number of *L. seymouri* cells x 10^{-4} (b) amount of pM8 DNA (ng). Probe was radiolabeled LINS 1 oligonucleotide (see Figure 2A). All DNAs were digested with *Hin*dIII and denatured prior to filtration onto the slot blot apparatus.

number of each repeat unit, we hybridized a genomic Southern blot with the 35 nt MX oligonucleotide. Under stringent hybridization conditions (hybridization between homologies of ≤28 bp would not be detected), the ratio of hybridization intensities was 10:7:3 for the 0.75, 1.1 and 5.9 kb *Hin*dIII restriction fragments (data not shown). Similar results were observed in blots where the entire 0.75 kb gene of pM4 was used as a probe (Figure 1B, lanes 1,2,5,12 and Figure 7B, lanes 5,6). Thus, this ratio represents the distribution of mini-exon sequences as well as mini-exon-associated sequences (defined as those present in the 0.75 kb gene) in the *L. seymouri* genome. Since there are a total of 425 copies of the mini-exon per cell, there are approximately 225 copies of the 0.75 kb gene family, 140 copies of the 1.1 kb family, and 60 copies within the 5.9 kb family.

Organization of mini-exon genes in *L.seymouri*

As several enzymes with single recognition sites within pM4 generate 0.75 and 1.1 kb fragments in genomic digests, the 0.75 and 1.1 kb genes could be

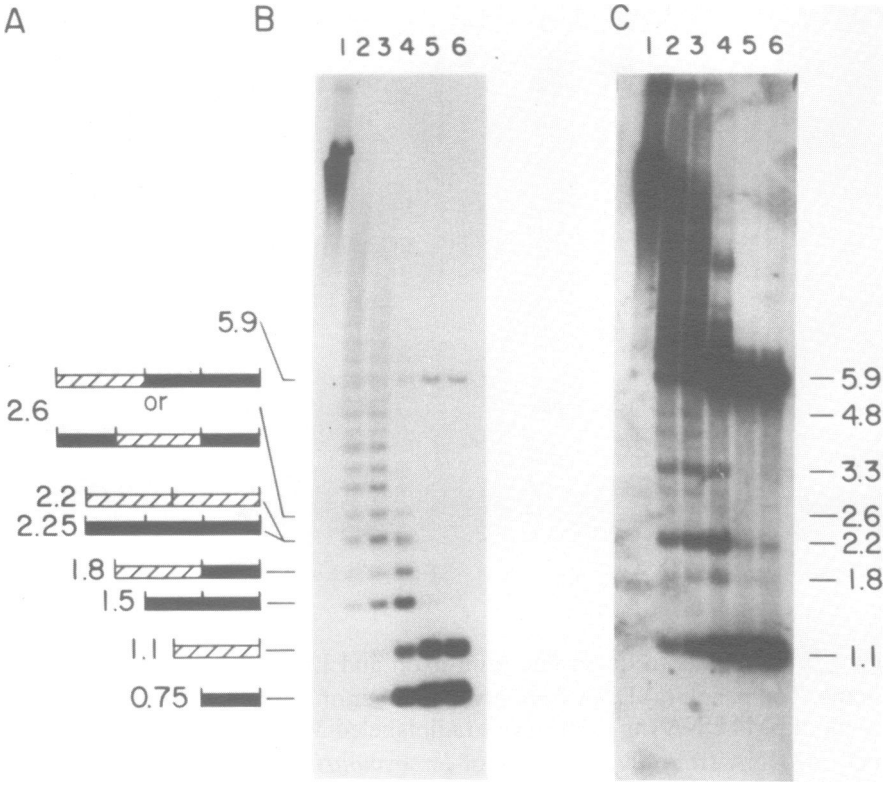


Figure 7. (A) Diagram of 0.75 and 1.1 kb gene arrangements that correspond to some of the hybridized fragments shown to the right. Black boxes represent the 0.75 kb genes and hatched boxes represent the 1.1 genes. The numbers represent sizes in kb. (B) Southern blot of a time course of *Hin*dIII-digestions of genomic DNA probed with radiolabelled pM4. Lanes 1,2,3,4,5,6 indicate digestion for 0,2,5,10,20,30 min respectively. Exposure was 2 days without intensifying screen. (C) Same as (B), except probe was pLINS 1. Exposure was 3 days with intensifying screen.

tandemly repeated in a head to tail fashion in one of several possible arrangements. To determine if the two gene types are arranged in individual or interspersed arrays, partial digestion products of *Hin*dIII-restricted genomic DNA were analysed (Figure 7). Panel B shows the hybridization pattern obtained when the blot was probed with the 0.75 kb gene present on clone pM4. Panel A depicts the hybridized fragments and the 0.75 and 1.1 kb gene organizations they represent. The presence of the 1.8 kb fragment indicates interspersion of 0.75 and 1.1 kb genes in the tandem array. Fragments of 1.5 and

2.2 kb demonstrate clustering of at least two copies of a 0.75 or 1.1 kb gene within an array. The 2.25 kb band, slightly above the 2.2 kb fragment, indicates clustering of three copies of 0.75 kb genes (the doublet is clear in the original autoradiogram). Larger clusters, of three, four, or five copies of either 1.1 or 0.75 kb genes, were consistent with the presence of hybridizing fragments of 3.0, 3.7, 4.0, 4.4 and 5.5 kb. Therefore, the tandem array consists of interspersed clusters of 0.75 and 1.1 kb genes. Each cluster probably contains from one to several copies of a single gene because partial digestion products such as the 1.8 and 2.6 kb fragments, which contain both genes and thus are 'junction' fragments, are readily discerned and therefore relatively abundant. Panel C shows that pLINS 1, a probe specific for the unique region of pM8, in relation to pM4, hybridized to 2.2 and 3.3 kb partial digestion products that most likely represent multimers of 1.1 kb genes. In addition, hybridization to 1.8, 2.6, 3.0, and 4.8 kb bands likely are equivalent to those in panel B and thus represent 0.75 plus 1.1 kb-containing species.

The organization of LINS 1-containing mini-exon genes was determined by Southern blot analysis using a LINS 1-specific 225 bp sequence (see Figure 2B) as a probe (Figure 5). Hybridization to genomic 1.1 kb fragments generated by enzymes that recognize a single site within the homologous regions of the 0.75 and 1.1 kb genes showed that the LINS 1-containing 1.1 kb mini-exon genes were included within the head to tail, tandemly arrayed 0.75 and 1.1 kb mini-exon genes (Figure 5, lanes 1,2,5,6,7,12). LINS 1-containing 1.1 kb mini-exon genes contain a unique *Hin*cII site within the LINS 1 sequence (Figure 2B). Therefore, detection by the LINS 1-specific probe of a 1.1 kb hybridized fragment in *Hin*cII -digested genomic DNA indicated that at least two LINS 1-containing mini-exon genes occur in tandem (Figure 7, lane 17) . In addition, the existence of a 1.1-0.75-1.1 head to tail gene arrangement in which both 1.1 kb genes contain LINS 1 was shown by hybridization of LINS 1 probe to a 1.9 kb *Hin*cII fragment. The presence of the 2.9 and 3.9 kb *Hin*cII fragments was consistent with an interspersed, clustered gene arrangement of mini-exon sequences in which 0.75 and 1.1 kb genes, with and without intact LINS 1 sequences, are present.

A comparison of LINS 1 and mini-exon hybridization to *Hin*dIII-digested DNA indicated that most, if not all, LINS 1 sequences were contained within mini-exon genes. However, in some genomic digests certain restriction fragments were detected only with the LINS 1 probe. These fragments likely represent a subset of sequences contained within the 5.9 kb mini-exon-containing *Hin*dIII species since they hybridized with an intensity similar to that

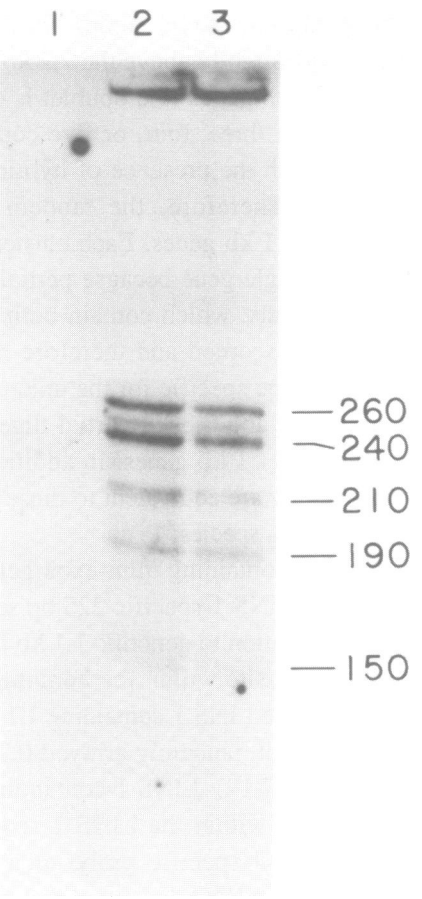


Figure 8. Northern blot of LINS 1 RNAs. 3 μg of poly $(A)^+$, 6μg total and poly $(A)^-$ *L. seymouri* RNA (lanes 1,2,and 3, respectively) were electrophoresed on a denaturing 8% polyacrylamide gel, transferred to Nytran and hybridized to a riboprobe transcribed from pLINS 1. This probe complements RNA that would be transcribed left to right on the pM8 schematic shown in Figure 2A. The sizes of the major and minor RNA species are indicated. Autoradiography was for 3 days with an intensifying screen.

of the 5.9 kb *Hin*dIII species (see Figure 5; bracketed bands).

Two restriction digests that did not yield 1.1 or 0.75 kb bands, because the enzymes did not cut within either of the repeat units, produced large (>20 kb) fragments and a pair of smaller ones that were all hybridized by both the mini-exon and LINS 1 probes (Figures 1B and 5, lanes 3 and 4). Hybridization to the smaller fragments corresponded in intensity to hybridization to the 5.9 kb

*Hin*dIII mini-exon-containing species for each of the probes. This indicated that these fragments may overlap with the 5.9 kb *Hin*dIII species. Therefore, the tandem clustered array of 1.1 and 0.75 kb genes is probably not extensively broken by the 5.9 kb *Hin*dIII species.

The 5.9 kb *Hin*dIII species and the other LINS 1-containing genes were all present on large *Bgl*II and *Cla*I fragments (Figure 5, lanes 8,9). This demonstrated that, like the LINS 1-containing 1.1 kb species, the 5.9 kb *Hin*dIII species is not randomly dispersed in the genome but clustered in one or several genomic locations.

Hybridization of the mini-exon to the 5.9 kb *Hin*dIII species indicated that either there are many copies of the mini-exon sequence within a single 5.9 kb fragment or that there are several copies of this species in the genome. Similarly, LINS 1 hybridization to this species does not distinguish between internal LINS 1 reiteration and a single LINS 1 copy within multiple copies of the 5.9 kb *Hin*dIII species. Cloning and sequence analysis of this *Hin*dIII fragment would be necessary to distinguish between the two possibilities.

Transcription of LINS 1

Four abundant and three minor transcripts were detected on Northern blots when the LINS 1-specific probe was hybridized to *L. seymouri* RNA (Figure 8). Experiments designed to detect large LINS 1-containing RNAs complementary to or equivalent to the medRNA coding strand indicated that the only stable LINS 1 RNAs are those shown in Figure 8. Thus, all LINS 1 RNAs are nonpolyadenlyated and are transcribed from left to right in the orientation of LINS 1 shown in Figure 2B. Analysis of ^{32}P-labeled nascent transcripts confirmed that transcription of LINS 1 is unidirectional (data not shown).

Because it is possible that transcription of LINS 1-containing 1.1 kb genes, such as that present in pM8, initiate at the mini-exon sequence and proceed through the LINS 1 sequence, we determined whether the detected LINS 1 RNAs contained mini-exon or other medRNA sequences. None of the LINS 1 RNAs shown in Figure 8 hybridized to the MX oligonucleotide or to riboprobes representing either strand of the entire *L. seymouri* insert in pM4. These hybridizations were performed under conditions where 26 bp of homology, specifically the AT- rich 26 bp of mini-exon sequence (nt 4 through 30), 5' to the LINS 1 sequence shown in Figure 2A, would be detected. Therefore, stable LINS 1 RNAs do not contain mini-exon sequences although it is unknown if this is the case for primary LINS 1 RNA transcripts.

DISCUSSION

L. seymouri has both uninterrupted and interrupted mini-exon genes. One example of each class was cloned and fully sequenced. The interrupted genes can contain at least three types of DNA insertions, each approximately 300 bp in length. One type of insertion, designated LINS 1, was extensively characterized as to its sequence, genomic abundance, location within a mini-exon gene, flanking sequences and transcriptional activity. The LINS 1 sequence contains no translation initiation sites, and the peptides deduced from all six reading frames share no homology with anything in the Genbank and Dayhoff databases. LINS 1 has no significant homology with any of the known *T. brucei* insertions, including RIME (29), INGI (30), TRS (31), and the approximately 400 bp of published SLACS/MAE sequence (8,9).

The RIME element was cloned in an interrupted ribosomal gene, where it contained a stretch of 14 A's, after the end of an ORF (open reading frame) that encoded a 160 amino acid protein, and was flanked by 7 bp duplications of the target DNA (29). TRS/INGI elements contain RIME sequences at their termini, a poly A tail at one end and an ORF with homology to reverse transcriptases (30-32). These elements are highly repeated, although only TRS/INGI are dispersed in the genome. SLACS or MAE sequences (8,9) are of low copy number, are only present within mini-exon gene arrays, are 5.5 to 7 kb long and, as shown in one report, contain a poly d(A) stretch distal to the interrupted medRNA gene. These facts, as well as target site duplication flanking SLACS and TRS/INGI, have resulted in the designation of these elements as retroposons (8,30,32). A several kb sequence that interrupts the mini-exon gene array in *C. fasciculata* also appears to be a retroposon (A. Gabriel, personal communication). A retroposon origin may help explain the presence of LINS 1 in mini-exon genes. Retroposons, which constitute a major class of transposable elements in eukaryotes, presumably result from reverse transcription of an RNA and subsequent insertion of the cDNA product into the genome by non-homologous transpositional recombination, which results in a short target site duplication due to DNA repair-synthesis at the closely spaced staggered breaks that accompany the largely uncharacterized mechanism of eukaroytic transposition.

The simplest class of retroposons are pseudogenes, which are formed by reintegration of a cDNA transcript of an mRNA. It is possible that LINS 1-containing mini-exon genes originated by the insertion of a pseudogene into the mini-exon gene array. Pseudogenes in trypanosomes would be expected to initially exist as cDNAs with a mini-exon at one end and a poly d(A) tract at the other, if they were derived from full length reverse transcription of any cellular

mRNA. Reverse transcription is potentially possible in trypanosomes, because TRS/INGI elements in *T. brucei* possess ORFs with homology to retroviral reverse transcriptases. The resultant cDNA could, after circularization, integrate preferentially into mini-exon genes by homologous recombination between mini-exon sequences.

LINS 1 and SLACS may be derived from the same original retroposon. They are present in relatively few copies per genome and are exclusively associated with medRNA genes. The SLACS-mini-exon gene boundary is at mini-exon base 12 and the LINS 1-mini-exon gene boundary is at mini-exon base 30. The size difference, as well as the short potential displaced poly d(A) stretch in LINS 1 (A_6TA_5, 26 nt from the 3' flanking duplication) and its lack of protein coding capacity, indicate that LINS 1 may be a degenerated retroposon. Partial homology between LINS 1 and a second, incompletely characterized, mini-exon insertion sequence that has a poly d(A) tract at one end, suggests that sequence divergence may have occurred after the original pseudogene was integrated into the genome and the interrupted mini-exon gene copy expanded. However, it is possible that the 5.9 kb *Hin*dIII fragment, which most likely contains a much larger inserted sequence in the mini-exon as well as LINS 1 sequences, may be analogous, homologous or closely related to the SLACS sequence found in *T. brucei*.

Repetitive elements in eukaryotes can be generated by several mechanisms: repeated transposition of a mobile element, the expansion of sequences by recombinational means such as unequal crossovers, gene conversion events, or a combination of these processes. Because the LINS 1 sequence is exclusively associated with mini-exon sequences, it appears not to be a mobile genetic element. A single insertion of a LINS 1 precursor element (or pseudogene) into a mini-exon gene and subsequent expansion of sequences by unequal crossing over could explain the interspersed clustered arrangement of the LINS 1-containing mini-exon genes.

Sequences 5' to the medRNA coding region were compared among eight trypanosomatids to try to identify common elements that may function as RNA polymerase recognition sites. Three short conserved regions were found. The presence of these regions suggests that RNA polymerase recognition of promoter regions in mini-exon genes may extend 265 bp upstream from the mini-exon sequence. However, none of these three regions have any homology with known consensus sequences used by RNA polymerases I, II or III in other eukaroytes. This may be expected because some of the unique aspects of trypanosome transcription may reflect the presence of trypanosome-specific

RNA polymerase factors and unique DNA recognition signals. Four levels of α-amanitin sensitivity can be distinguished during transcription in isolated *T. brucei* nuclei. By comparison to other eukaroytic systems, the intermediate levels of α-amanitin resistance exhibited by medRNA, as well as the lack of polyadenylation, suggests that these genes are transcribed by an RNA polymerase III-type enzyme. Polymerase III genes generally contain internal regions that serve as enzyme recognition sites (33). These conserved sequences (A and B boxes) have no similarity to internal medRNA sequences.

If the sequences conserved between the 0.75 and 1.1 kb repeat units serve to initiate and terminate medRNA transcription then we would expect to see a med-LINS 1 RNA of 379 nt (293+86) at about 1/10th the intensity on Northern blots as the 86 nt medRNA. medRNA-LINS 1 transcripts were expected because, in addition to the sequence arrangement of LINS 1 and medRNA in pM8, we have shown that most if not all LINS 1 sequences are associated with mini-exon sequences in the 5.9 kb *Hin*dIII and 1.1 kb *Hin*dIII fragments. Analysis of stable and nascent RNA showed that the LINS 1 sequence is transcribed in the same direction as a medRNA-LINS 1 transcript would be, and that RNAs containing LINS 1 sequences are not inherently unstable. The surprising absence of stable medRNA-LINS 1 composite RNAs indicates that either a LINS 1-containing medRNA is even more unstable than is medRNA (6 min half-life in *T. brucei,* ref. 34), or that, if an internal sequence is required for medRNA gene recognition by an RNA polymerase III-like enzyme, interruption by LINS 1 could destroy, or move relative to upstream elements, these polymerase binding signals. It is also possible that LINS 1 RNAs are not encoded by the 1.1 kb gene, but by LINS 1 inserts present in the 5.9 kb *Hin*dIII genomic sequences. Another possibility is that primary LINS 1 transcripts possess the 5' 30 nt of the mini-exon, and this region is removed during RNA maturation, perhaps being spliced onto some mRNAs. Potential splice sequences are present close to the mini-exon-LINS 1 junction.

The relatively small variation among all sequenced mini-exons would suggest that the mini-exon sequences of trypanosomatids may be highly constrained for some as yet unknown reason. Thus, it may be unremarkable that the mini-exon sequences of *L. seymouri, L. enriettii* (7) and *C. fasciculata* (6,22) are identical. Alternatively, this identity may reflect a close evolutionary relationship among these organisms.

In summary, we have identified and sequenced the genes that encode the medRNA of *L. seymouri.* The medRNA is similar in size and sequence to the medRNA characterized in other trypanosomatids, and certain sequence motifs

are conserved within the regions 5' and 3' to the medRNAs. The role of mini-exon, medRNA and flanking sequences in RNA transcription and processing will become apparent only after genetic analysis.

ACKNOWLEDGEMENTS

We thank Dr. MonaLisa Mojumdar for isolating clone pM4, Angela F. de Amorim for technical assistance, and C. Graham Clark for useful discussions and critical reading of the manuscript. We thank Dr. Abram Gabriel for communicating data prior to publication. This work was supported by NIH award AI 21729.

*To whom correspondence should be sent

REFERENCES

1. Van der Ploeg, L.H.T., Cornelissen, A.W.C.A., Michels, P.A.M. and Borst, P. (1984) Cell 39, 213-221.
2. Krause, M. and Hirsh, D. (1987) Cell 49, 753-761.
3. Cook, G.A. and Donelson, J.E. (1987) Mol. Biochem. Parasit. 25, 113-122.
4. 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.
5. Nelson, R.G., Parsons, M., Barr, P.J., Stuart, K., Silkirk, M. and Agabian, N. (1983) Cell 34, 901-909.
6. Muhich, M.L., Hughes, D.E., Simpson, A.M. and Simpson, L. (1987) Nucl. Acids Res. 15, 3141-3153.
7. Miller, S., Landfear, S. and Wirth, D. (1986) Nucl. Acids Res. 14, 7341-7360.
8. Aksoy, S., Lalor, T.M., Martin, J., Van der Ploeg, L.H.T. and Richards, F.F. (1987) EMBO J. 6, 3819-3826.
9. Carrington, M., Roditi, I. and Williams, R.O. (1987) Nucl. Acids Res. 15, 10179-10198.
10. Bone, G.T. and Steinert, M. (1956) Nature 178, 308.
11. Cully, D.F., Ip, H.S. and Cross, G.A.M. (1985) Cell 42, 173-182.
12. Feinberg, A.P. and. Vogelstein, B. (1983) Anal. Biochem. 132, 6-13.
13. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, NY.
14. Bellofatto, V. and Cross, G.A.M. (1988) Nucl. Acids Res. 16, in press.
15. Grunstein, M. and. Hogness, D.S. (1975) Proc. Natl. Acad. Sci. USA 72, 3961-3965.
16. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
17. Tabor, S. and Richardson, C.C. (1987) Proc. Natl. Acad. Sci. USA 84, 4767-4771.
18. Martinez, H.M., Katzung, B. and Farrah, T. (1984) Sequence analysis program manual, UCSF, San Fransisco, California.
19. Freistadt, M.S., Cross, G.A.M., Branch, A.D. and Robertson, H.D. (1987) Nucl. Acids Res. 15, 9861-9879.

20. Freistadt, M.S., Cross, G.A.M. and Robertson, H.D. (1988) J. Biol. Chem. 263, in press.
21. Perry, K.L., Watkins,K.P. and Agabian, N. (1987) Proc. Natl. Acad. Sci. USA 84, 8190-8194.
22. Gabriel, A., Sisodia, S.S. and Cleveland, D.W. (1987) J. Biol. Chem. 262, 16192-16199.
23. Kooter, J.M., DeLange, T. and Borst, P. (1984) EMBO J. 3, 2387-2392.
24. Hasan, G., Turner, M.J. and Cordingley, J.S. (1984) Gene 27, 75-86.
25. Lenardo, M.J., Dorfman, D.M., Reddy, L.V. and Donelson, J.E. (1985) Gene 35, 131-141.
26. DeLange, T., Berkvens, T.M., Veerman, H.J.G., Carlos, A., Frasch, C., Barry, J.D. and Borst, P. (1984) Nucl. Acids Res. 12, 4431-4443.
27. Milhausen, M., Nelson, R.G., Sather, S., Selkirk, M. and Agabian, N. (1984) Cell 38, 721-729.
28. Dorfman, D.M. and Donelson, J.E. (1984) Nucl. Acids. Res. 12, 4907-4920.
29. Hasan, G., Turner, M.J. and. Cordingley, J.S. (1984) Cell 37, 333-341.
30. Kimmel, B.E., Ole-Moiyoi, O.K. and. Young, J.R. (1987) Mol. Cell. Biol. 7, 1465-1475.
31. Murphy, N.B., Pays, A., Tebabi, P., Coquelet, H., Guyaux, M., Steinert, M. and Pays, E. (1987) J. Mol. Biol. 195, 855-871.
32. Pays, E. and Murphy, N. (1987) J. Mol. Biol. 197, 147-148.
33. Sakonju, S. and Brown, D.D. (1982) Cell 31, 395- 405.
34. Laird, P.W., Zoomerdijk, J.C.B.M., de Korte, D. and Borst, P. (1987) EMBO J. 6, 1055-1062.