Volume 14 Number 18 1986 Nucleic Acids Research

# Cloning and characterization of a Leishmania gene encoding a RNA spliced leader sequence

Samuel I.Miller, Scott M.Landfear and Dyann F.Wirth

Department of Tropical Public Health, Harvard School of Public Health, Boston, MA, USA

Received 11 May 1986; Revised and Accepted 8 August 1986

ABSTRACT

Recent studies on Leishmania enriettii tubulin mRNAs revealed a 35 nucleotide addition to their 5' end. The gene that codes for this 35 nucleotide leader sequence has now been cloned and sequenced. In the Leishmania genome, the spliced leader gene exists as a tandem repeat of 438 bases. There are approximately 150 copies of this gene comprising 0.1% of the parasite genome. This gene codes for a 85 nucleotide transcript that contains the spliced leader at its 5' end. The 35 nucleotide sequence and the regions immediately 5' and 3' to it are highly conserved across trypanosomatids. We have detected a RNA molecule that is a putative by-product of the processing reaction in which the 35 nucleotide spliced leader has been transferred to mRNA. We suggest that this molecule is the remnant of the spliced leader transcript after removal of the 35 nucleotide spliced leader.

## INTRODUCTION

Leishmania are parasitic protozoa of the family Trypanosomatidae (1). These parasites have two developmentally distinct stages. Amastigotes are non-motile forms that live within mammalian host macrophages. Promastigotes are motile, flagellated forms that develop from amastigotes after ingestion by sandflies.

Previous work in this laboratory has focused on the mechanisms responsible for the control of expression of Leishmania α- and β- tubulin genes (2). When the mature messenger RNAs coding for α- and β- tubulin genes were analyzed using S1 nuclease mapping and primer extension techniques, they were found to have at their 5' ends a 35 nucleotide sequence which was not encoded contiguously with the remainder of the tubulin mRNAs (3).

Recent work in African trypanosomes (4-12) as well as other trypanosomatids (8, 9, 11, 13) has demonstrated a common 35 nucleotide spliced leader sequence at the 5' end of messenger RNAs. The gene coding for this spliced leader (SL) is tandemly repeated in the genome of trypanosomes (5, 7, 12) and can be present on a different chromosome from an expression linked copy of a variant surface glycoprotein (VSG) gene (10). The trypanosome

© IRL Press Limited, Oxford, England.

spliced leader gene encodes a transcript of approximately 137 nucleotides that contains the 35 nucleotide spliced leader at its 5' end. This 35 base sequence is also present at the 5' end of trypanosome calmodulin (14), VSG (4-9) and tubulin messenger RNAs (15). Analogous genes and their small RNA transcripts have been identified in the trypanosomatids, Trypanosoma cruzi, Trypanosoma vivax, and Leptomonas collosoma (9, 11, 13).

We now report the isolation of a similar gene in Leishmania enriettii and the identification of its transcript. This gene exists in high copy number as a tandem repeat within the genome and codes for a transcript of approximately 85 nucleotides. At the 5' end of this transcript is the 35 nucleotide sequence that is present on the 5' end of Leishmania α- and β- tubulin mRNAs (Landfear, Miller, and Wirth; Molecular and Biochemical Parasitology in press).

We also report the identification of a putative by-product of the reaction in which the 35 nucleotide spliced leader has been removed. We suggest that this RNA molecule is the remnant of the spliced leader transcript after transfer of the 35 nucleotide spliced leader to messenger RNA. Elucidation of the complete structure of this RNA processing by-product should help define the mechanism of spliced leader addition to messenger RNA.

## MATERIALS AND METHODS

### Isolation of organisms and nucleic acids

Promastigotes and amastigotes of L. enriettii were grown and purified as previously described (2, 16). DNA and RNA were purified by phenol-chloroform extraction as detailed elsewhere (16). Guanidinium purified RNA was prepared by vortexing a pellet of L. enriettii cells in 10 volumes of solution containing 6 M guanidinium thiocyanate, 5 mM sodium citrate pH 7.0, 0.1 M β-mercaptoethanol and 0.5% sarkosyl (17). This solution was layered over a 5.7 M cesium chloride solution and spun in a Beckman SW 50.1 rotor at 35,000 revolutions/minute for 16 hours at 20°C. Polyadenylated RNA was purified by binding total nucleic acid to a column of oligo dT cellulose type 3 (Collaborative Research) in a solution consisting of 400 mM NaAc, 5 mM EDTA and 0.1% SDS; then the bound RNA was eluted in a solution of 0.01 M NaAc, 1 mM EDTA (18). The fraction which did not bind to oligo dT cellulose was referred to as nonpolyadenylated RNA.

### Isolation of clone p11-28

High molecular weight genomic DNA from L. enriettii was digested to completion with the restriction enzyme Rsa I (New England Biolabs, Beverly, Mass.). This DNA was size-fractionated on a 0.8% agarose gel run in 50 mM

Tris-acetate pH 8.0, 1 mM EDTA. DNA in the size range of 320 to 600 base pairs was recovered by the glass powder extraction method (19). This size fractionated DNA was cloned into the Cla I site of the plasmid vector pBR322. Prior to ligation, the plasmid DNA was digested with Cla I, and incubated with the Klenow fragment of DNA polymerase I and deoxynucleotides (dGTP and dCTP) under conditions described by the manufacturer to fill in the 5' overhang at the Cla I site. The plasmid DNA was also treated with calf intestinal phosphatase prior to ligation (17). The ligation mixture was used to transform (17) E. coli HB101 cells. Transformants were plated onto agar containing 100 µg/ml ampicillin and incubated overnight at 37°C. Colonies were transferred to nitrocellulose filters, lysed in sodium hydroxide and neutralized (17). The filters were baked for 2 hours at 70°C and then hybridized at 42°C in 50% formamide, 5X SSC (1X SSC is 150 mM NaCl, 15 mM Na citrate), 20 mM $NaPO_4$ pH 7.0, 10X Denhardts and 100 µg/ml denatured salmon sperm DNA. The filters were probed with 400 µg of high molecular weight L. enriettii chromosomal DNA radiolabeled by nick-translation, washed three times for 30 minutes at 50°C in 0.1 X SSC, 0.5% SDS and exposed to film. Twenty colonies that gave strong signals (indicating a highly repeated sequence within the genome) were then colony purified, and plasmid DNA preparations were made using the alkaline lysis method (17). The plasmids were digested with Eco RI and Hind III restriction enzymes and run in a 0.8% agarose gel. The DNA was transferred to nitrocellulose by the method of Southern (17) and hybridized at reduced stringency (42°C in 30% formamide, 5X SSC, 20 mM NaPO4 pH 7.0, 10X Denhardts and 100 µg/ml denatured salmon sperm DNA) to a cloned purified gene fragment (radiolabeled by nick-translation) which codes for the spliced leader sequence of Trypanosoma brucei [Clone pDC-11-305 courtesy Drs. Doris Culley and George Cross at Rockefeller University, New York, New York]. One positive clone, p11-28, with an insert size of 438 base pairs was further analyzed.

DNA Sequencing

Complete sequence from both strands was obtained by the dideoxy chain termination method of Sanger (20) as modified by the New England Biolabs for the M13 sequencing system. The p11-28 plasmid was digested with restriction enzymes to generate DNA fragments for ligation and cloning into the M13 phage polylinker. Restriction enzyme sites Alu I (position 389) and Hae III (position 127) of the p11-28 insert were used in conjunction with the restriction enzyme sites Eco RI, Hind III, and Alu I of the plasmid vector pBR322 to make DNA fragments which allowed sequencing of the majority of both DNA strands. To complete the entire sequence of both DNA strands, specific

Bal 31 nuclease deleted clones were constructed. The p11-28 plasmid (100 µg) was digested with either Eco RI or Hind III restriction enzymes, incubated with 1.8 units of Bal 31 nuclease for 20 minutes, end filled with the Klenow fragment of DNA polymerase I, and then digested with the restriction enzyme that had not been used previously. The resulting fragments were then gel purified using the glass powder extraction method (19) and ligated into mp8 and mp9 sequencing vectors. Sequencing reactions were analyzed in 6% and 8% polyacrylamide 8 M urea sequencing gels that were electrophoresed in buffer containing 2 mM EDTA, 100 mM Tris-borate, pH 8.3. Gels were fixed in a solution of 10% acetic acid and 10% methanol, dried, and exposed without an intensifying screen. The length of sequence obtained from specific clones is shown in Figure 1A.

Analysis of small molecular RNAs by blot hybridization

The RNA was resolved by electrophoresis in a 20% polyacrylamide gel containing 8 M urea in buffer containing 2 mM EDTA, 100 mM Tris-borate, pH 8.3. The RNA in the gel was transferred electrophoretically to Gene Screen Plus hybridization transfer membrane (Dupont) at 4 volts/cm for 12 hours in a solution of 6 mM NaAc, 0.3 mM EDTA, 12 mM Tris-acetate pH 7.5. Filters were air dried and incubated for 6 hr at 42°C in a solution containing 50% formamide, 1% SDS, 1 M NaCl, and 10% dextran sulfate. Then denatured salmon sperm DNA (250 µg/ml and radiolabeled plasmid DNA (5 ng/ml) were added and the incubation continued for another 16 hours. Filters so treated were washed two times in 2X SSC at 26°C for 5 minutes, two times in 2X SSC, 1% SDS at 60°C for 30 minutes, and two times in 0.1 X SSC at 26°C for thirty minutes (1X SSC is 150 mM NaCl, 15 mM Na citrate). Filters were air dried and exposed to Kodak XAR-5 film with an intensifying screen.

S1 Nuclease Mapping

For S1 nuclease mapping, a M13 clone containing the entire p11-28 sense strand was chosen. Approximately 150 ng of this phage DNA was hybridized to the New England Biolabs sequencing primer and second strand synthesis accomplished using the Klenow fragment of DNA polymerase I. The probe was uniformly labeled at low specific activity. The reaction mix contained 20 µCi of all four $\alpha$ $^{32}$P-dNTPs (specific activity 800 Ci/mM), and an excess of unlabeled dNTPs (35 µM). The product of this reaction was then digested with restriction enzymes, Eco RI and Hind III, and run in a 1.5% agarose gel. The uniformly labeled insert was then gel-purified by the glass powder extraction method.

Three reactions were prepared, one containing DNA alone, one containing

DNA and S1 nuclease and one containing DNA, RNA and S1 nuclease. RNA and equal amounts of radiolabeled DNA were dried under vacuum, and dissolved in 25 µl of 80% formamide, 0.4M NaCl, 40 mM Pipes (pH 6.4), and 1 mM EDTA. Samples were heated to 80°C for 10 minutes and incubated at 47° for 3 hours. The DNA-RNA hybrids were diluted into 250 µl ice cold buffer containing 30 mM NaAc (pH 4.6), 250 mM NaCl, 1 mM $ZnSO_4$, 5% glycerol, 20 µg/ml denatured salmon sperm DNA and quick frozen. The hybrids were incubated with or without S1 nuclease (1000 U/ml BRL) at 45°C for 30 minutes. Digestion was terminated by addition of EDTA to a final concentration of 5 mM. Calf liver transfer RNA (10 µg) was added as carrier and the mixture was phenol/chloroform extracted and ethanol precipitated. Samples were heated to 95°C in 50% deionized formamide and analyzed by electrophoresis in an 8% polyacrylamide 8 M urea gel as in DNA sequencing.

Primer extension using an oligonucleotide primer and reverse transcriptase

An oligonucleotide primer complementary to nucleotides 24-38 of the p11-28 insert (see Figure 1A) was hybridized to RNA in 80% formamide, 0.4 M NaCl, 40 mM Pipes (pH 6.4), and 1mM EDTA for 2 hours at 60°C (5 µg non-polyadenylated promastigote phenol/chloroform purified RNA, 16 µg guanidinium purified promastigote total RNA). The hybrids were precipitated by the addition of two volumes of ethanol and dissolved in 25 µl containing 50 mM Tris-HCl, pH 8.3, 140 mM KCl, 7 mM $MgCl_2$, 10 mM dithiothreitol, RNasin 2.4 u/ul (New England Biolabs), 100 µM dCTP, 100µM dTTP, 100µM dGTP and 2.5µM dATP with 25 µCi $\alpha$ $^{32}$P-dATP. Reverse transcriptase 5 units (Life Sciences Inc, St. Petersburg, FL) was added and the mixture was incubated at 42°C, for 15 minutes, and then chased for 15 minutes with the addition of an excess (85 µM) of unlabeled dNTPs at 42°C. The reaction was terminated by the addition of an equal volume of deionized formamide containing 10 mM NaOH. Reactions were heated to 95°C, and loaded in an 8% polyacrylamide 8M urea gel. Gels were fixed and dried as for sequencing.

Primer extension sequencing

The reaction mix was the same as above for primer extension. At time 0 the reaction mixture was aliquoted to four separate reactions each containing one of the four 2', 3'-dideoxynucleoside triphosphates as chain terminators in a 1:1 ratio to the deoxynucleoside triphosphates. The dNTP concentrations were 2.5 uM for dATP and 100 uM for all others. Sixty-four µCi $\alpha$ $^{32}$P-dATP were added to the reaction (800 Ci/mM). After 15 minutes incubation at 42°C, the reaction was chased with the addition of an excess of unlabeled dNTPs (85 µM). The reaction was terminated by the addition of an equal volume of

deionized formamide containing 10 mM NaOH. The samples were analyzed by electrophoresis as described above for DNA sequencing reactions.

## RESULTS

### Isolation of the spliced leader gene

Initially a Leishmania enriettii sequence homologous to the Trypanosoma brucei spliced leader was identified by Southern blot analysis of restriction enzyme cut DNA. The size class of Rsa I cut L. enriettii DNA that hybridized to the T. brucei spliced leader gene was ligated into pBR322. The resulting recombinant plasmids were transformed into HB101 cells, and this size selected library was screened. Clones containing highly repeated Leishmania DNA sequences were identified using nick-translated total genomic DNA from L. enriettii. A similar approach was used to identify the T. brucei spliced leader gene (5). Then plasmid DNAs of the positive clones from this screen were isolated, digested with restriction enzymes to free the L. enrietti inserts and hybridized with the T. brucei clone for the spliced leader gene. Two plasmids, p11-28 and p16-4, had inserts homologous to the T. brucei spliced leader gene, and clone p11-28 was analyzed further.

To confirm that this clone encoded the spliced leader of L. enriettii, the DNA sequence of the p11-28 insert was determined. The 438 base pair sequence of the p11-28 insert is shown in Figure 1A. Recently, we (Landfear, Miller, and Wirth; Molecular and Biochemical Parasitology in press) have demonstrated that the first 35 nucleotides at the 5' end of α- and β- tubulin mRNAs are identical to bases 1-9 and 413-438 of the p11-28 insert as shown in figure 1A. Since the gene was cloned from a genomic Rsa I digest, the cloned DNA is cut in the spliced leader sequence.

A comparison of this sequence with other known trypanosomatid spliced leader genes shows that these sequences have homologies within the 35 nucleotide spliced leader sequence and within the regions immediately 5' and 3' to this leader sequence. The bases surrounding the 35 nucleotide spliced leader sequence are identical across species despite the variability in the spliced leader sequence itself. In addition, all spliced leader genes have a run of thymidine nucleotides near the 3' termination of transcription (figure 1B), similar to the run of 9 thymidine nucleotides at position 66-74 (figure 1A) in the p11-28 insert. The L. enriettii spliced leader gene has no other significant homology with the published sequences of the T. brucei and T. cruzi spliced leader genes. The L. collosoma and L. enriettii spliced leader

genes are more similar in having additional homology in the transcribed bases 3' to the spliced leader. Specifically in a region beginning 12 nucleotides 3' to the spliced leader (figure 1B), eleven of the next twelve nucleotides are identical in the gene sequences of L. enriettii and L. collosoma.

The spliced leader genes are arranged in a tandem repeat with some restriction site polymorphism

Southern blot analysis of different restriction enzyme digests of Leishmania enriettii were performed to analyze the arrangement of p11-28 sequences within the genome. The results in Figure 2A demonstrate that the leader sequence genes possess a single recognition site for Rsa I. This restriction site is within the 35 nucleotide leader sequence and is highly conserved across species (Figure 1B). Partial digestion of L. enriettii chromosomal DNA with the restriction enzyme Rsa I revealed a ladder of hybridizing bands separated by 438 bases in length. This demonstrates that the spliced leader gene exists as a tandem repeat within the parasite genome (figure 2A). Quantitative Southern blotting was performed to determine the genome copy number of the spliced leader gene. Known amounts of Rsa I digested chromosomal DNA and p11-28 plasmid insert were size fractionated on an agarose gel, blotted onto nitrocellulose and probed with the p11-28 (spliced leader) insert. The film was then scanned using a densitometer. This demonstrated that the spliced leader gene comprises approximately 0.1% of the parasite genome and that it exists in approximately 150 copies per haploid genome (data not shown).

Figure 2B shows that digestion of chromosomal DNA with Pst I (lane 3), an enzyme which does not cut within the repeat unit, produces spliced leader gene fragments of high molecular weight (greater than 23 kilobases). Furthermore, the Pst I digestion does not reveal any additional lower molecular weight bands, suggesting that there are no orphon copies of this gene. This is in contrast to the chromosomal arrangement of the spliced leader genes in T. brucei in which orphon gene copies have been well documented (21). The results in figure 2B also shows that restriction site polymorphism occurs within the repeat. Specifically, the enzyme Hinf I (lane 1) cuts rarely and releases fragments of different sizes; in contrast Hae III (lane 4) occasionally cuts more than one time within the repeat. The p11-28 clone appears to be typical of many of the repeats in that it has a single Hae III site. Other enzymes (Nar I, Hinc II, Alu I, Acc I) that cut once within most units of the repeat release hybridizing fragments of the same size as the Rsa I digest (data not shown).

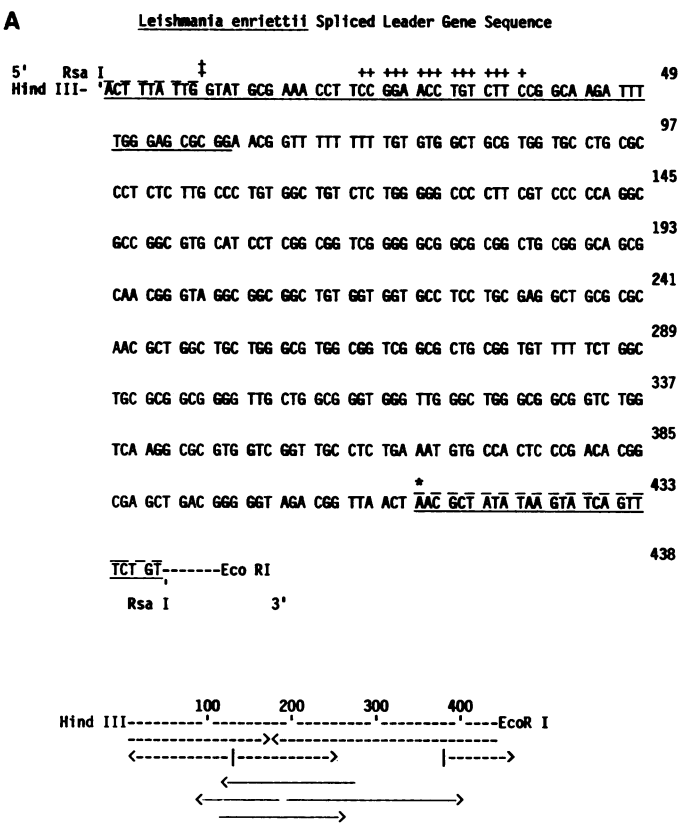



Figure 1 A: The 438 nucleotide genomic sequence of the sense strand of the Leishmania enriettii spliced leader gene as cloned into the plasmid pBR322 is shown above. The Hind III and Eco RI restriction enzyme sites of pBR322 are indicated in their respective 5' and 3' positions. The Rsa I restriction enzyme site in genomic DNA that was destroyed by the construction of this plasmid is indicated at the ends of the L. enriettii spliced leader gene sequence. The 35 nucleotide spliced leader sequence transferred to messenger RNA is indicated by a dashed line above those nucleotides. The asterisk (*) above nucleotide 413 indicates the 5' end of the transcript as well as the 5'

end of the spliced leader. The double cross (‡) indicates the 3' end of spliced leader. The single crosses (+) indicate the sequence homologous to the oligonucleotide used in primer extension. The transcribed region is underlined. The two restriction enzyme sites used in sequencing are an Alu I site at position 389 and an Hae III site at position 127. The schematic figure below the sequence indicates specific M13 phage clones that were sequenced. The numbered line indicates the gene ;the Eco RI and Hind III restriction enzyme sites of the plasmid pBR322 are indicated at their respective positions relative to the spliced leader insert. The dashed lines represent sequence read from clones generated by restriction enzyme digestion and the solid lines represent sequence read from clones generated by Bal 31 nuclease digestion. The arrows represent the strand from which the sequence was read. Note that there is complete overlap of the entire sequence as read from both DNA strands. B: Comparison of genomic sequences of spliced leader genes (5,9,11). The underlined nucleotides correspond to the nucleotide sequence present on the 5' end of messenger RNA. The asterisks indicate nucleotides conserved in all species. The crosses indicate nucleotides conserved between L. collosoma and L. enriettii in the transcribed region 3' to the leader sequence present on mRNAs. The sequences correspond to sense strands. All transcripts end within ten nucleotides 5' to the runs of thymidine bases at the 3' end of sequences shown. All transcripts analyzed begin at the 5' end of the underlined leader sequence present on mRNA.

## The spliced leader gene hybridizes to many messenger RNAs and codes for a discrete transcript

To identify a transcript of the Leishmania enriettii spliced leader gene, parasite RNA was subjected to Northern blot analysis. Figure 3A demonstrates that the p11-28 clone hybridizes to a discrete RNA of less than 310 nucleotides in total (lanes 2 and 3) and non-polyadenylated (lane 5) RNA, but not in polyadenylated RNA (lane 4). Additionally, in polyadenylated RNA, the leader sequence gene hybridizes to a broad size range of RNAs (lane 4). This suggests that the Leishmania enriettii spliced leader gene contains a sequence common to many polyadenylated RNAs, as has been demonstrated for the T. brucei spliced leader gene (5,7,12).

To determine the size of the small, major transcript more accurately, RNA was analyzed on polyacrylamide gels, electrophoretically transferred to Gene Screen Plus hybridization filter, and probed with the p11-28 plasmid. Figure 3B shows the result of this analysis of Leishmania enriettii RNA. One can see that the RNA band hybridizing to the spliced leader gene corresponds to 85 nucleotides when compared to double stranded DNA markers. The band runs larger than a sequenced yeast transfer RNA of 76 nucleotides (data not shown, transfer RNA courtesy Dr. L. Gherke, M.I.T.).

Results in figure 3A (lanes 2 and 3), and 3B (lane 2), also show that the spliced leader gene transcript is present in greater relative quantity in amastigote RNA compared to promastigote RNA. This is a reproducible result

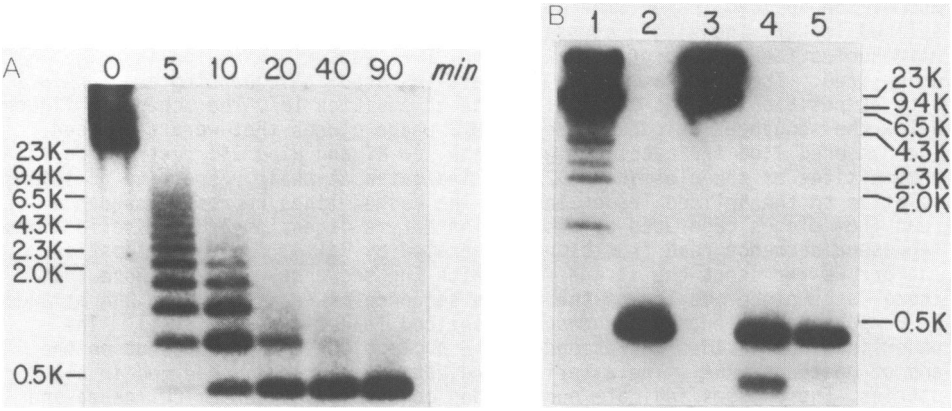



Figure 2.A: Genomic Southern blot of an Rsa I partial digest of Leishmania DNA probed with the spliced leader gene p11-28. Ten µg of L. enriettii chromosomal DNA from promastigotes was digested with 50 units of Rsa I restriction enzyme. At various time points, EDTA to a final concentration of 50 mM was added to aliquots of the reaction containing 2 µg of chromosomal DNA, and the mixture was extracted with phenol/chloroform. The DNA from each time point (5', 10', 20', 40', 90') was analyzed by electrophoresis in a 0.8% agarose gel at 85 volts for two hours. The DNA was transferred to nitrocellulose by the method of Southern (17). The nitrocellulose papers were baked at 70°C for two hours and then prehybridized for 2 hours at 42°C in 50% formamide, 5X SSC, 20 mM $NaPO_4$ (pH 7.0), 10X Denhardt's and 100 µg/ml salmon sperm DNA and hybridized for 16 hours with radiolabelled p11-28 plasmid, prepared by nick-translation, in the same solution. The filter was washed extensively at 50°C in 0.1X SSC, 0.5% SDS and exposed to film for three hours. Molecular weight markers of a λ Hind III digest are indicated in kilobases.
B: Genomic Southern blot of Leishmania DNA probed with the spliced leader clone p11-28. The experimental conditions were the same as Figure 2A except all restriction digests were 2 hours in length and monitered for complete digestion. All lanes contain 2 µg of chromosomal DNA. Lane 1, is a Hinf I digest; lane 2, a Rsa I digest; lane 3, a Pst I digest; lane 4, a Hae III digest; lane 5, an Alu I digest. Molecular weight markers of a λ Hind III digest are indicated in kilobases.

and experiments are in progress to discover the explanation of this result. Furthermore, our preparation of guanidinium extracted RNA (30 µg per lane) shows that the transcript is present in low abundance compared to phenol/chloroform extracted RNA (10 µg per lane, figure 3B). The low intensity hybridizing bands of lower molecular weight seen in the amastigote lane in figure 3B could represent RNA species derived from the spliced leader transcript, but no characterization of these bands has been possible because of their low abundance.

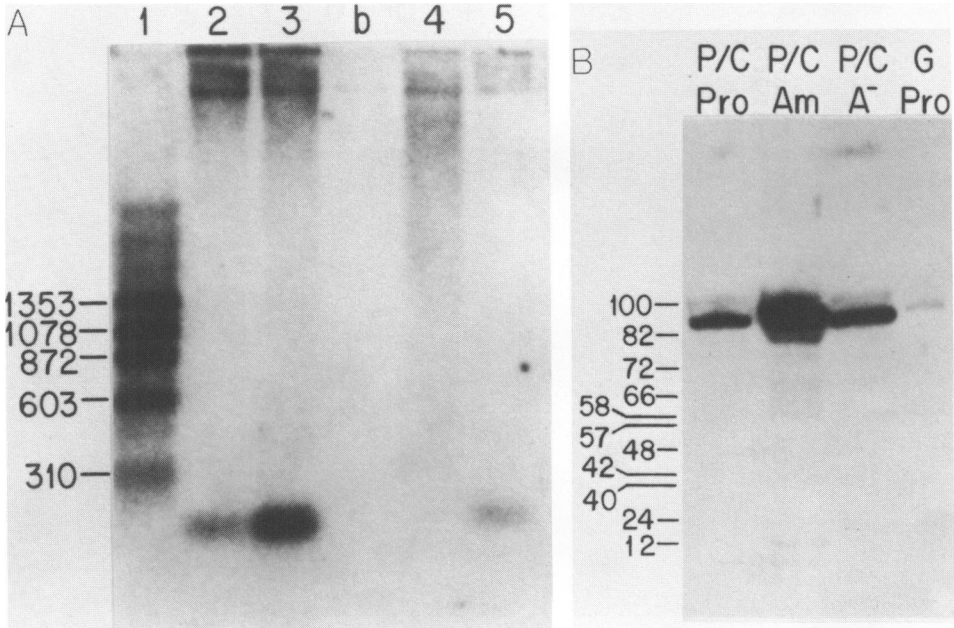
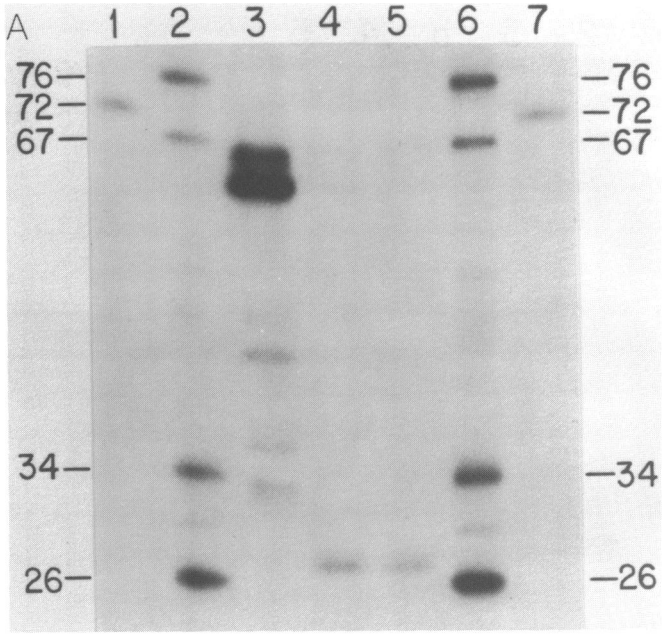
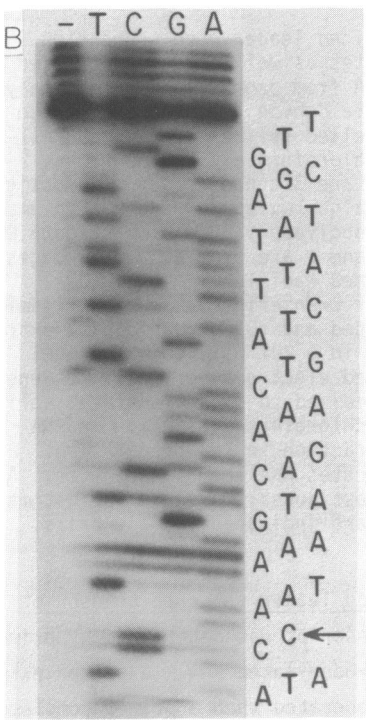


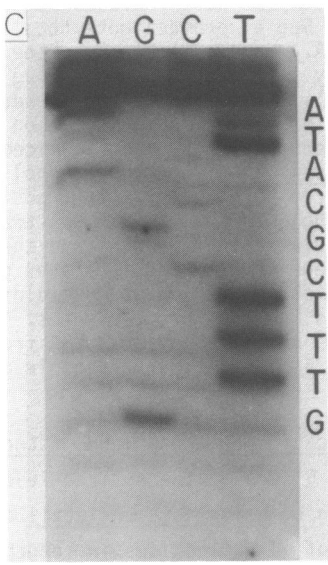

Figure 3. A: Northern analysis of the spliced leader gene transcript. Total RNA from promastigotes and amastigotes as well as polyadenylated RNA from amastigotes and nonpolyadenylated RNA from promastigotes was analyzed by electrophoresis in a 1.5% agarose, 6% formaldehyde gel, transferred to nitrocellulose and hybridized to radiolabelled spliced leader gene p11-28. All RNA samples were prepared by phenol/chloroform extraction. Lane 2 contains 5µg of promastigote total RNA. Lane 3 contains 5µg of amastigote total RNA. Lane b is blank. Lane 4 contains 1µg of polyadenylated amastigote RNA (poly A+), Lane 5 contains 5µg of nonopolyadenylated promastigote RNA (poly A-). Molecular weight markers in lane 1 are from a Hae III digest of ØX174. The hybridization solution contained radiolabeled ØX174 DNA. B: Northern analysis of the spliced leader transcript in a polyacrylamide gel. L. enriettii RNA 10 µg/lane phenol extracted and 30 µg guanidinium extracted was electrophoretically size fractionated in a 20% polyacrylamide gel containing 8M urea. The RNA was transferred electrophoretically to Gene Screen Plus hybridization membrane and then hybridized to nick-translated spliced leader gene p11-28. PC indicates phenol/chloroform extraction in RNA preparation. G indicates guanidinium extraction in RNA preparation. Pro indicates RNA from promastigotes. AM indicates RNA from amastigotes. A- indicates nonpolyadenylated RNA from promastigotes. Molecular weight markers of a ØX174 Hae III, Hinf I double digest are indicated in base pairs.

## Identification of the 5' end of the spliced leader sequence transcript

The 5' end of the spliced leader transcript was determined by primer extension of Leishmania enriettii non-polyadenylated RNA. A primer extension product of 64 nucleotides in length was generated when a primer complementary


A
1 2 3 4 5 6 7
76
72
67
34
26
76
72
67
34
26


B
- T C G A
G A T T A C A C G A A C A
T G A T T T A A A T A A C T A
T C T A C G A G A A T C


C
A G C T
A T A C G C T T T G

to nucleotides 24-38 (figure 1A) of the p11-28 insert was used (figure 4A, lane 3). This result locates the 5' end of the spliced leader transcript at the start of the 35 nucleotide spliced leader sequence. Primer extension sequencing of the non-polyadenylated RNA, that served as template to the primer, was performed to define the exact nucleotide at which transcription initiates. The results in figure 4B show that the 5' end of the RNA terminates with a strong stop at the first base 5' to the 35 nucleotide spliced leader sequence. Four strong stops were observed immediately after this initial strong stop. These strong stops may indicate alternate start sites for transcription or putative processing just 5' to the spliced leader in the region of the gene conserved across species.

S1 nuclease mapping of the 3' end of the spliced leader transcript

From polyacrylamide gel analysis, the spliced leader transcript is approximately 85 nucleotides in length. Primer extension sequencing positions the 5' end of the transcript at nucleotide 413 in the p11-28 insert (fig. 1A). High resolution S1 mapping was performed to map the 3' end of this transcript. Uniformly labeled insert from the p11-28 clone was hybridized to the non-polyadenylated fraction of L. enriettii RNA and digested with S1 nuclease under conditions emprically determined in which RNA:DNA hybrids of greater

Figure 4. A: Primer extension of Leishmania enriettii RNA.
Autoradiogram of an 8% acrylamide/8M urea gel containing primer extension products generated when 120 ng of an oligonucleotide primer complementary to nucleotides 24-38 (fig 1A) of the spliced leader gene p11-28 was hybridized to 5 µg of promastigote nonpolyadenylated RNA prepared by phenol/chloroform extraction (Lane 3) or 16 µg (Lane 4) and 8 µg (Lane 5) of promastigote total RNA prepared by guanidinium extraction and extended using reverse transcriptase in the presence of $\alpha$ $^{32}$P-dATP and unlabeled nucleotides. Lanes 1 and 7 contain ØX Hae III markers that have been radiolabeled with gamma-$^{32}$P-dATP. Lanes 2 and 6 contain a Msp I digest of pBR322 that has been end filled at the 3' end with $\alpha$ $^{32}$P-dNTPs. The numbers indicate the lengths of molecular weight markers in nucleotides.
B: Primer extension sequencing of phenol/chloroform extracted RNA.
After hybridization to the same oligonucleotide primer as in 4A, the reaction mixture was aliquoted to 4 separate reactions containing one of four 2', 3'-dideoxynucleoside triphosphates in a 1:1 ratio to deoxynucleoside triphosphates and extended using reverse transriptase. The products of these reactions were run on a 10% acrylamide/8M urea gel. Lane (-) contains no dideoxynucleotide. Lanes A ,G ,C ,and T correspond to the dideoxynucleotide present in the reaction mix either adenine, guanosine, cytosine, or thymidine. The arrow indicates the 3' end of the spliced leader transferred to messenger RNA.
C: Primer extension sequencing of guanidinium extracted RNA.
The reactions were identical to those in 4B. Total promastigote guanidinium extracted RNA (16µg) was used. The products of the reactions were run on a 20% acrylamide/8M urea gel. Lanes A,G,C,and T correspond to the dideoxynucleotide present as in B.

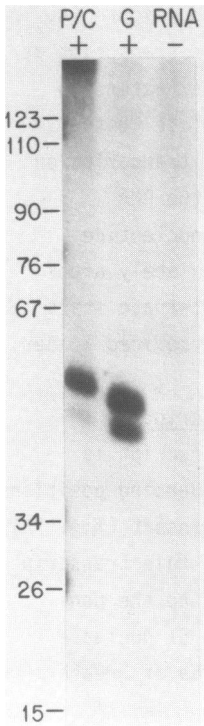



Figure 5. S1 nuclease mapping of the spliced leader gene transcripts. Autoradiogram of an 8% acrylamide/8M urea gel containing S1 nuclease protected fragments. Uniformly radiolabeled coding strand of p11-28 was constructed using an single stranded M13 clone containing the sense strand of p11-28. This was hybridized at 47°C to 5 µg phenol/chloroform extracted nonpolyadenylated promastigote RNA (P/C +), 16 µg guanidinium extracted total promastigote RNA (G +) and digested with S1 nuclease. A control sample to which no RNA was added (RNA -) was also digested with S1 nuclease. The protected fragments corresponded to the 3' end of the leader gene transcript as, at these conditions, the 26 bases at one end of p11-28 (Figure 1A) which correspond to the first 26 bases of the 5' end of the transcript, are not protected. The numbers indicate the lengths in nucleotides of an endfilled Msp I digest of pBR322.

than 26 base pairs were detected. Based on the predicted transcript size, we expected a protected fragment of 55-60 nucleotides representing the 3' end of this fragment. As can be seen in figure 5, a fragment of this length (approximately 59 nucleotides) is indeed protected. In addition, a second fragment of 50 nucleotides is also protected. This smaller fragment was enriched relative to the larger fragment in the RNA that had been extracted in guanidinium.

Identification of a processing product

To determine the structure of the RNA corresponding to the smaller S1 nuclease protected fragment, we conducted primer extension studies on the guanidinium extracted RNA. As can be seen in figure 4A (lanes 4 and 5), a smaller primer extension product is found which is exactly 35 nucleotides shorter than the full length transcript. One interesting possibility is that the smaller RNA molecule containing spliced leader genomic sequence is the by-product remaining after the 35 nucleotide sequence is donated from the spliced leader transcript to the 5' end of messenger RNA. The size of the smaller protected fragment (50 bases) in the S1 mapping experiment (figure 5)

is also consistent with this hypothesis. The radiolabeled probe used in mapping contained 9 bases of the 35 nucleotide leader sequence contiguous with the remainder of the spliced leader transcript. Hence, if the full length transcript protects a DNA fragment of 59 bases, then the putative by-product should protect a fragment of 50 bases.

To test this hypothesis RNA was sequenced by primer extension. As shown in figure 4C, the smaller primer extension product is indeed the product left after the 35 nucleotide spliced leader is removed. The sequence is identical to the spliced leader transcript, but its 5' end terminates at the splice junction site. Thus, we appear to have identified one of the by-products of the spliced leader reaction.

One puzzling result is that there is no full-length primer extension product found when guanidinium RNA is used (figure 4A, lanes 4 and 5). Utilizing guanidinium extracted RNA we are able to detect the full length spliced leader transcript, albeit at reduced levels, using Northern blot analysis (figure 4B) and S1 nuclease mapping (figure 5). We have no explanation for this apparent discrepancy between the S1 nuclease mapping and primer extension results, but it is also found in other preparations of guanidinium extracted RNA. It may be due to the relative abundance of the putative processing product versus the primary 85 nucleotide transcript or some secondary structure not formed in the phenol/chloroform extracted RNA.

## DISCUSSION

Recent studies on trypanosomatids, particularly African trypanosomes, have documented that these organisms synthesize messenger RNA by a discontinuous transcription mechanism (4-12). Specifically, messenger RNAs contain a common 35 nucleotide leader sequence (4) that is transcribed separately from the body of the mRNA (5,9,12). The leader sequence is encoded by a high copy number, tandemly repeated gene (5,7,12). This gene is transcribed to yield a small RNA (85-140 nucleotides in different species) that contains the 35 nucleotide spliced leader at its 5' end (5,9,11,12,13). All messenger RNAs examined to date from trypanosomatids contain a spliced leader. In African trypanosomes, the spliced leader gene can be transcribed from a different chromosome than the remainder of variant surface glycoprotein mRNAs (10). Biochemical evidence in African trypanosomes suggests that the leader sequence initial transcript is capped and that its polymerase has α-amanitin sensitivity similar to the polymerase involved in the transcription of tubulin genes (22).

We have cloned and analyzed the Leishmania enriettii spliced leader gene in order to understand the mechanism of processing and transcription of Leishmania tubulin messenger RNAs. Our studies have shown that the leader sequence gene exists as a tandem repeat with some restriction site polymorphism. In contrast to Trypanosoma brucei, we find no evidence in L. enriettii that copies of the spliced leader gene exist separately from a tandem repeat. The spliced leader gene comprises approximately 0.1% of the parasite genome and exists in approximately 150 copies per haploid genome. It hybridizes in a smear pattern to polyadenylated RNAs in Northern blots, suggesting that it has a common sequence with many messenger RNAs. Its primary transcript contains the leader sequence on its 5' end, is approximately 85 bases in length, and is not polyadenylated. Primer extension sequencing of the 5' end of the transcript accurately defines that end of the molecule to the nucleotide; the 3' end of the transcript is approximately 85 nucleotides downstream from this site. The initial spliced leader transcript is similar in size to the L. collosoma transcript and, like those identified in other species, ends near a run of thymidine bases (figure 1B).

Comparison among trypanosomatid spliced leader genes shows conservation of the leader sequence, its Rsa I restriction enzyme site and the regions immediately 5' and 3' to the leader sequence. There is no other significant homology among all trypanosomatids within the remainder of the gene except for a run of thymidine nucleotides seen in all species 3' to the end of transcription (figure 1B). There is conservation between L. collosoma and L. enriettii in a small transcribed region of the gene 3' to the spliced leader sequence (figure 1B). Conservation of the leader sequence and the regions surrounding the leader sequence in all trypanosomatids yet analyzed suggest that their primary structure is important. Our primer extension sequencing of the spliced leader initial transcript demonstrates that the 4 conserved bases 5' to the 35 nucleotide spliced leader create strong stops. These bases may represent alternative transcription start sites or processing products.

Using S1 nuclease mapping and primer extension techniques, we have identified a second RNA species that contains spliced leader genomic sequence. This molecule could be a by-product of the reaction in which the spliced leader is transferred to mRNA, as its 5' end is the remnant of the spliced leader transcript. By primer extension sequencing we define the 5' end of this putative processing product exactly at the dinucleotide GU which occurs at the junction of the 35 nucleotide spliced leader sequence with the remainder of the transcript. All spliced leader transcripts examined to date

contain the dinucleotide GU immediately 3' to the 35 nucleotide spliced leader sequence transferred to mRNA (fig 1B). These bases conform to a eucaryotic splice junction sequence at the 5' end of introns (23). Similarly all mRNA encoding genes examined from trypanosomatids to date, [tubulin genes from Leishmania enriettii (Landfear, Miller, Wirth; Molecular and Biochemical Parasitology in press), variant surface glycoprotein (5, 7, 11), tubulin (15), and calmodulin (14) genes from African trypanosomes], contain the dinucleotide AG immediately 5' to the junction for addition of the spliced leader to mRNA . These bases conform to a eucaryotic splice junction sequence at the 3' end of introns (23). These consensus splice junction sequences suggest that RNA splicing analogous to other eukaryotic splicing is involved in the transfer of the spliced leader from its transcript to mRNA.

Since mRNAs in trypanosomatids can be assembled from transcription products of two unlinked genes, the synthesis of these mRNAs must involve at least two transcriptional events. Two kinds of splicing, cis and trans, (figure 6) have been proposed to explain the RNA processing after discontinuous transcription (5,9). In one cis-splicing model, the initial leader transcript could be used as a primer to transcribe structural genes. After primed transcription and a splicing event to remove a putative intron, mature mRNA is generated. Primed transcription could also occur by a mechanism in which the 3' end of the spliced leader transcript is first cleaved and then the 35 nucleotide spliced leader is used to prime transcription. A polymerase priming mechanism is used by influenza virus, where a host RNA is used as a primer to transcribe viral genes (24). Alternatively, the spliced leader transcript might be ligated to the primary transcript of a mRNA encoding gene; this ligated precursor would then be processed by cis-splicing to remove an intervening sequence and generate mature mRNA.

Another model that has been proposed is that of a bimolecular trans-splicing reaction (5,9). In this model the spliced leader gene and the mRNA encoding gene are initially transcribed as two separate molecules; then a bimolecular splicing reaction would produce mRNA and remnants of the initial transcripts. Recently trans-splicing of two separate synthetic RNAs of two different species has been demonstrated in vitro using mammalian cell splicing extracts (25,26).

Our identification of a RNA molecule containing sequence of the spliced leader gene but whose 5' end terminates with the first base 3' to the 35 nucleotide leader sequence suggests we have found a by-product of the reaction

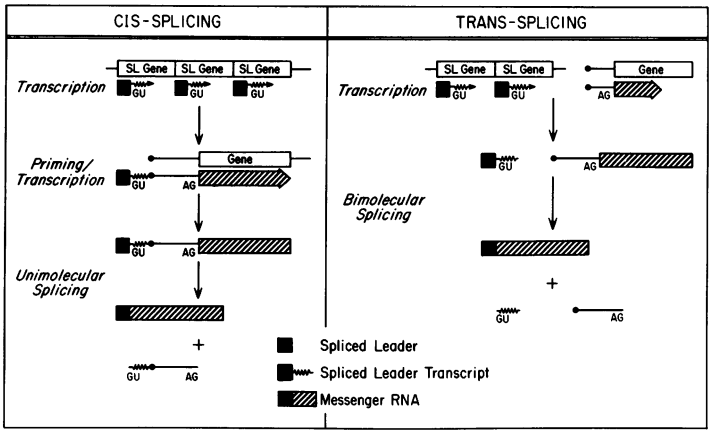



Figure 6. Models of RNA processing after discontinuous transcription (5,9). In cis-splicing the spliced leader transcript could prime transcription of a gene resulting in a precursor with an intron. Unimolecular splicing produces a single RNA by-product containing genomic sequence upstream to the AG consensus splice site and the remainder of the spliced leader transcript. Alternatively in cis-splicing a ligation reaction after two unlinked transcription events could result in a unimolecular splicing event. In trans-splicing independent transcription events followed by bimolecular splicing produce the two RNA by-products shown above and mature mRNA. One RNA by-product contains spliced leader genomic sequence and a separate by-product contains transcribed genomic sequence from the gene processed to mRNA. Primed transcription after removal of the 3' end of the spliced leader transcript would also result in a RNA processing product comprising the free 3' end of the spliced leader transcript. GU represents the nucleotides of the consensus eucaryotic splice junction found in the spliced leader transcript immediately 3' to the spliced leader. AG represents the nucleotides of the consensus eucaryotic splice junction found immediately 5' to the splice junction in all trypanosomatid mRNA genes examined to date.

in which the spliced leader sequence has been cleaved from its primary transcript. One surprising result is that we can not identify this molecule in blot hybridization in preparations of guanidinium extracted RNA that are enriched in these putative by-products (fig. 3B). One possible explanation for this result is that the 3' end of this putative product is different in each separate mRNA, as would be predicted by the cis-splicing models. In this case, the total pool of processing products would have the same 5' end, as detected by techniques of S1 mapping and primer extension, but due to different 3' ends, they would be heterogeneous in size. Their abundance could be too low to detect in blot hybridization experiments, and, if detected, they

would appear as a smear rather than a discrete band. The smaller hybridizing smear and bands in the amastigote lane in Figure 3B could represent these putative by-products; alternatively, they could simply be degradation products of the spliced leader transcript.

In contrast, if the putative RNA by-product only contains the remaining 3' end of the spliced leader transcript, as predicted by the trans splicing model, it would be approximately 50 nucleotides in length. We have been unable to detect a discrete band of this length in blot hybridizations (fig. 3B). However, if the molecule we identified has a lariat structure similar to that found in eukaryotic introns after splicing, it might not migrate as a linear RNA of 50 nucleotides (27). With this in mind, we have examined these RNA molecules in polyacrylamide gels at different concentrations in two dimensions (data not shown); however, we do not observe any aberrantly migrating species that would suggest a lariat structure.

Regardless of its structure further characterization of this remnant of the spliced leader primary transcript and putative by-product of splicing should help distinguish between the models of spliced leader addition (fig 6). If its 3' end is identical to the 3' end of the initial spliced leader transcript and contains no other gene sequence, this would favor either a bimolecular trans splicing mechanism or a splicing event involving the removal of the 3' end of the spliced leader transcript before the leader sequence was used to prime transcription. If the 3' end of this putative by-product contains transcribed genomic sequence upstream from the consensus AG splice site of a mRNA encoding gene, then this would favor one of the cis-splicing mechanisms. Determination of the structure of the 3' end of this putative splicing by-product should help to distinguish between these models.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institutes of Health (AI21365-01A1), John D. and Catherine T. MacArthur Foundation and the Rockefeller Foundation awarded to D.F.W. and Dr. John David. S.I.M. is supported by a National Institutes of Health, National Research Services Award, 5 F32 AI07179-02. S.M.L. acknowledges a fellowship from the Medical Foundation. D.F.W. is a Burroughs Wellcome Fund Scholar in Molecular Parasitology. We want to thank Drs. Doris Culley and George Cross for the use of the pDC-11-305 clone as well as Dr. Thomas Unnasch and Dr. Cynthia French for helpful discussion.

Abbreviations: EDTA, ethylenediamine-tetracetic acid; SDS, sodium dodecyl sulfate; TRIS-HCL, TRIS(hydroxy methyl) aminomethane-chloride; NaCl, sodium chloride; NaAC, sodium acetate; DNA, deoxy ribonucleic acid; RNA, ribonucleic acid; $NaPO_4$, sodium phosphate; PIPES, piperazine-N-N'-bis[2-ethanesulfonic acid]; $ZnSO_4$, zinc sulfate; dCTP, deoxycytidine triphosphate; dTTP, deoxythymidine triphosphate; dGTP, deoxyguanine triphosphate; dATP, deoxyadenine triphosphate; dNTP, deoxynucleoside triphosphate.

REFERENCES


1. Zuckerman, A. and Lainson, R. (1979) in Parasitic Protozoa, Kreier, J.P. Ed., Vol. I, pp57-133, Academic Press, New York.
2. Landfear, S.M. and Wirth, D.F. (1984) Nature 309, 716-717.
3. Landfear, S.M. and Wirth, D.F. (1985) Mol. Bioch. Parasitol. 15, 61-82.
4. Boothroyd, J.C. and Cross, G.A.M. (1982) Gene 20, 281-289.
5. Campbell, D.A., Thornton, D.A. and Boothroyd, J.C. (1984) Nature 311, 350-355.
6. Van der Ploeg, L.H.T., Liu, A.Y.C., Michels, P.A.M., DeLange T., Borst, P., Majunder, K., Weber H., Veeneman G.H., and Van Boom,J. (1982) Nuc. Acids Research 12, 3591-3604.
7. Parsons, M., Nelson, R.G., Watkins, K.P. and Agabian, N. (1984) Cell 38, 309-316.
8. Kooter, J.M., DeLange, T., and Borst, P. (1984) EMBO J. 10, 2387-2392.
9. Milhausen, M., Nelson, R.G., Sather, S., Selkirk, M. and Agabian, N. (1984) Cell 38, 721-729.
10. Van der Ploeg, L.H.T., Cornelissen, A.W.C.A., Michels, P.A.M. and Borst, P. (1984) Cell 39, 213-221.
11. DeLange, T., Berkiens, T.M., Veerman, H.J.G., Carlos, A., Frasch, C., Barry, J.D., and Borst, P. (1984) Nuc. Acids Research 12, 4431-4443.
12. 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.
13. Nelson, R.G., Parsons, M.B., Selkirk, M., Newport G., Barr, P.S., and Agabian, N. (1984) Nature 308, 665-667.
14. Tschudi, C., Young A.S., Ruben, L., Patton, C.L. and Richards, F.F. (1985) Proc. Natl. Acad. Sci. USA 82, 3998-4002.
15. Sather, S. and Agabian, N. (1985) Proc. Natl. Acad. Sci. USA 82, 5695-5699.
16. Landfear, S.M., McMahon-Pratt, D. and Wirth, D.F. (1983) Mol. Cell. Biol. 3, 1070-1076.
17. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning :A Laboratory Manual, Cold Spring Harbor Laboratory, New York.
18. Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69, 1408-1412.
19. Vogelstein, B. and Gillespie, D. (1979) Proc. Natl. Acad. Sci, USA 76, 615-619.
20. Sanger, F., and Coulson, A.R., (1978) FEBS Letter 87, 107-110.
21. Nelson, R.G., Parsons, M., Barr P.J., Stuart K., Selkirk, M., and Agabian N. (1983) Cell 34, 901-909.
22. Laird, P.W., Kooter, J.M., Loosbrook, N. and Borst, P. (1985) Nucleic Acids Research 13, 4253-4266.
23. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) Cell 38, 317-331.
24. Krug, R.M. (1985) Cell 41, 651-652.
25. Konarska, M.M., Padgett, R.A. and Sharp, P.A. (1985) Cell 42, 165-171.
26. Solnick, D. (1985) Cell 42, 157-164.
27. Grabowski, P.J., Padgett, R.A., and Sharp, P.A. (1984) Cell 37, 415-427.