Nuclease mapping and DNA sequence analysis of transcripts from the dihydrofolate reductase-thymidylate synthase (R) region of *Leishmania major*

Geoffrey M.Kapler+, Kang Zhang and Stephen M.Beverley*
Department of Biological Chemistry and Molecular Pharmacology, 250 Longwood Ave, Harvard Medical School, Boston, MA 02115, USA

Received April 20, 1990; Revised and Accepted June 8, 1990 EMBL accession nos X51733, X51734

ABSTRACT

Trypanosomatid protozoan parasites utilize a number of nonstandard mechanisms in expressing their genes. To probe these phenomena in a genetically accessible system, we have mapped termini of eight transcripts arising from the amplified R region including the DHFR-TS gene of methotrexate-resistant *Leishmania major*. Poly(A)+ RNAs transcribed from the DHFR-TS-coding strand exhibit features similar to those observed around other trypanosomatid protein-coding genes. These include close spacing, the presence of a trans-spliced miniexon on the 5′ termini, heterogeneity at both 5′ and 3′ ends, and in some cases S1 nuclease protection of intertranscript regions. Other than the splice acceptor site, no consensus sequence elements associated with either 5′ or 3′ ends were detected, although polydinucleotide tracts tended to be near inter-transcript regions. Two poly(A)+ RNAs transcribed from the opposite strand of the upstream flanking regions lacked the miniexon. Sequencing of DNA encoding the overlapping 1.7 kb opposite strand transcripts (one bearing and one lacking the miniexon, both found on polysomes) revealed no reading frames likely to encode proteins, suggesting that at least some of these RNAs could be nonfunctional by-products of RNA processing.

INTRODUCTION

Trypanosomatid protozoa employ a variety of novel processes to express their genes, suggesting that the rules governing transcription are different from those of other eukaryotes (1–8). Transcription of trypanosomatid protein-coding genes yields a series of polyadenylated RNAs encoded in close proximity within the DNA, separated by short intertranscript regions. These RNAs bear the trans-spliced 39 nucleotide miniexon at their 5′ end. In contrast to most other species intertranscript regions are generally transcribed at a rate similar to that of the regions encoding adjacent mature transcripts. This suggested the possibility that transcription occurs in a 'polycistronic' fashion, and that processing generates multiple mature mRNAs (7; 9–13). Consistent with the polycistronic model, conserved DNA sequences which commonly mark promoters in other organisms are usually absent in the immediate upstream regions of trypanosomatid genes (reviewed in 8). For some protein-coding genes, other transcripts arising from the putative polycistronic domain appear to lack significant open reading frames (14). These transcripts may be non-functional by-products analogous to introns, although these RNAs occur on polysomes (14; 15). Many questions remain concerning the specific mechanisms and results of polycistronic transcription in the loci studied thus far, and it is unknown whether polycistronic transcription is a universal feature of all trypanosomatid genes. Genetic tests of polycistronic vs. monocistronic transcription has not yet been obtained due to the very recent emergence of DNA transfection techniques (16–19).

We have utilized drug-resistant lines of *Leishmania major* for addressing a variety of molecular and biochemical questions in trypanosomatids. Methotrexate (MTX) resistant mutants often exhibit selective gene amplification of two separate regions of DNA, the R and H regions. The R region, encodes the bifunctional enzyme dihydrofolate reductase-thymidylate synthase (DHFR-TS; 20,21), a normally single-copy gene which is amplified in the R1000-3 line as a 30 kilobase (kb) extrachromosomal DNA (22; Fig. 1). Previous analysis of the DNA sequence immediately upstream of the DHFR-TS mRNA did not identify conserved sequences present in other DHFR and TS genes which are known to promote transcription and bind trans-acting factors (23–27). Analysis of the transcripts of the R region revealed that approximately 95% of it is transcribed into 9–10 stable polyadenylated RNAs (Fig. 1), all of which are associated with polysomes (15). These include the 'downstream' RNAs, which are transcribed from the same DNA strand and overlap one another to varying extents, and two examples of opposite-strand 'antisense' RNAs localized to the

* To whom correspondence should be addressed

+ Present address: Department of Microbiology and Immunology, University of California, San Francisco, CA, USA

divergent RNA subregion, upstream of DHFR-TS (Fig. 1). R region transcripts were identical in structure in the R1000-3 and wild-type lines, indicating that all cis-acting elements necessary for transcription initiation and RNA processing reside within the amplified region (15; 19). In this report we have characterized transcripts of the amplified R region in more detail by nuclease protection studies and DNA sequencing.

METHODS AND MATERIALS

Parasites

Promastigotes (the insect stage of this digenetic parasite) of the R1000-3 line of *Leishmania major* (28) were grown in M199 medium as described (19) containing 1 mM MTX. Large scale cultures were grown in medium containing 100 μM MTX.

RNA isolation, S1 nuclease mapping and primer extension

Total and poly(A)$^+$ RNAs were prepared as described (29,30). S1 nuclease mapping was performed using 2 μg of total cellular poly(A)$^+$ RNA (23). Primer extension was performed using specific oligonucleotide primers (at least 20 bases) whose 5′ ends were located as described in the text and 2 μg of total cellular poly(A)$^+$ RNA (15). Product sizes were estimated from multiple experiments using three different sets of molecular weight markers, and in all cases the standard deviation was less than 5 nucleotides.

Polymerase chain reaction (PCR) amplification of cDNA

cDNA synthesized from RNA-specific primers and 0.2 μg poly(A)$^+$ RNA was subjected to PCR amplification as previously described (15,31), using the cDNA synthesis primer and either miniexon primer A (GGGAATTCGGATCC/AACGCTATATAAGTTATCAG), which contains nucleotides (nt) 1 to 19 of the *L. major* miniexon (J. Miller, submitted for publication) and a 14 nt 5′ extension, or miniexon primer B (TCAGTTTCTGTACTTTATTG) which contains nt 16–35 of the miniexon.

DNA sequence analysis

DNAs were cloned into M13mp18 and/or M13mp19 vectors (32) and sequenced by the dideoxy method with modified T7 polymerase (Sequenase; U.S. Biochemical, Cleveland, Ohio). The complete sequence of both DNA strands was obtained. The sequence of the Sal I fragment containing the *L. major* DHFR-TS sequence (including Fig. 5B) was assigned accession number X51733 and the sequence of the divergent region (Fig. 5A) was assigned X51734.

RESULTS

RNA termini mapping

Downstream and DHFR-TS subregions. Northern blot mapping of the R region indicated that the subregion 3′ to the DHFR-TS mRNA encodes at least five RNAs collectively termed the downstream RNAs (3.2, 2.8, 2.5 and 2.3 kb), which are all transcribed from the same strand as DHFR-TS and overlap to varying extents (Fig. 1; 15); the 5′ ends of the 3.2 and 2.8 kb downstream RNAs mapped to a 0.6 kb SalI$_a$-SphI fragment of this region (Fig. 2B; 15). S1 nuclease analysis of this region produced a predominant protected fragment, mapping the 5′ end of the 3.2 and 2.8 kb downstream RNAs 405 bp to the right of the SalI$_a$ site; 3 additional minor 5′ ends were mapped 280, 385 and 430 bp to the right of this site (Fig. 2B). Since Northern

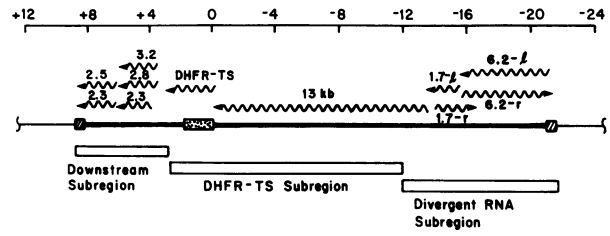


Figure 1. Overview of organization of the R region. The map of the wild-type chromosomal R region is shown; the heavy line denotes the segment which is amplified in the R1000–3 line by joining of the sites marked as hatched boxes. The stippled box indicates the DHFR-TS coding region. Coordinates above the map (in kb) indicate the position relative to the DHFR-TS start codon, increasing in the direction of DHFR-TS transcription. RNAs encoded are shown above the DNA map (15). For presentation the R region has been divided into arbitrary sub-regions as shown (downstream, DHFR-TS and divergent RNA).

blot hybridization with this fragment indicated that the 3.2 and 2.8 kb downstream RNAs were equally abundant (15), we presume that both RNAs utilize the major 5′ terminus identified here.

Primer extension of the downstream RNAs with an oligonucleotide whose 5′ end was located 39 bases to the right of the SalI$_a$ site (Fig. 2A) revealed two products with apparent 5′ ends mapping 318 and 438 bp to the right of the SalI$_a$ site (Fig. 2B), with the former of these products being the major species. These ends map 30–40 bp further upstream of minor and major S1 products (Fig. 2B, vertical dashed lines), suggesting that these RNAs contain miniexons on their 5′ ends. To confirm this, cDNA generated with the downstream RNA primer was subjected to PCR amplification with this primer in combination with either of two miniexon primers, A and B, which contain sequences from the 5′ and 3′ halves of the miniexon, respectively. If the miniexon were present, PCR amplification products would be present and be 30 nt larger when amplified with miniexon primer A than with primer B (miniexon primer A contains a 14 base 5′ extension in addition to the miniexon sequences). Aplification of cDNA synthesized from a downstream RNA-specific primer in conjunction with the A primer produced a major 385 bp product, while amplification with the B primer produced a major 360 bp product (Fig. 3; the size was determined more accurately from gels containing less DNA, not shown); other products may arise from alternative 5′ ends bearing miniexons. The PCR products were specific, since they hybridized to a downstream RNA-specific probe and their appearance was dependent on the presence of both primers (data not shown). The sizes estimated for these PCR products were slightly smaller than expected, based on the S1 and primer extension products discussed above. We attribute this discrepancy to differences in gel electrophoresis conditions, since a similar discrepancy was observed in our analysis of the DHFR-TS mRNA (19) for which the cDNA sequence had been determined. Thus, these data suggest that the 3.2 and 2.8 kb downstream RNAs have the miniexon at their 5′ end.

S1 nuclease mapping of the 3′ end of the DHFR-TS mRNA (Fig. 2A) revealed a predominant protected fragment terminating 560 bp to the left of the PvuII$_a$ site, in addition to less abundant 3′ ends mapping 470, 590 and 710 bp to the left of this site (Fig. 2B). A small amount of full length protection was also observed. The 560 bp site agrees well with that of a previously sequenced cDNA (20).

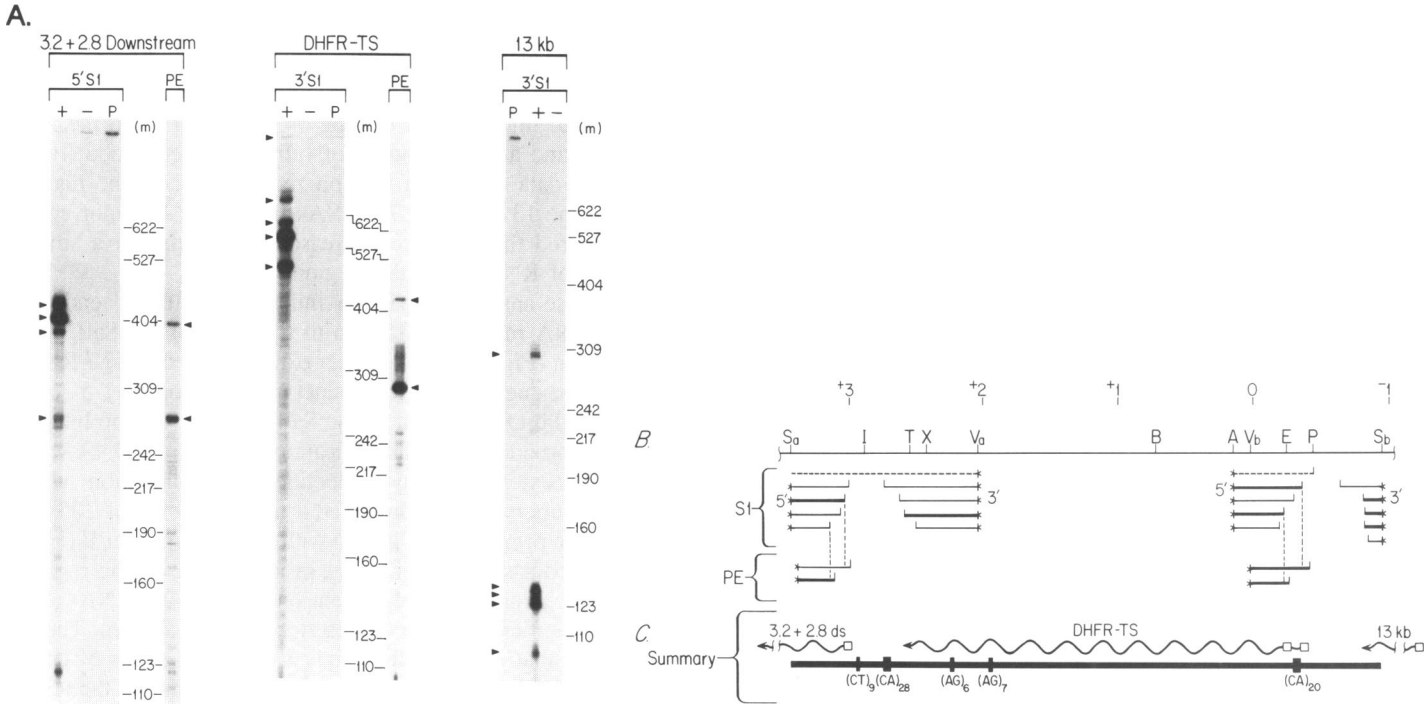


Figure 2. S1 nuclease protection and primer extension analyses of RNA termini flanking the DHFR-TS mRNA. Panel A: 1) S1 nuclease mapping of the 5′ end of the 3.2 and 2.8 kb downstream RNAs (lanes labeled 3.2+2.8 downstream, 5′S1), employing the 2.1 kb $SalI_a$-XhoI fragment (map position +3.5 to +1.4) labeled on the 5′ end of the SalI site (Fig. 2B); 2) Primer extension of the 2.8 and 3.2 kb downstream RNAs with an oligonucleotide primer labeled on the 5′ end 39 nt to the right of the $SalI_a$ site (lane labeled 3.2+2.8 downstream, PE); 3) S1 nuclease mapping of the 3′ end of the DHFR-TS mRNA (lanes labeled DHFR-TS, 3′ S1), using the 1.4 kb $SalI_a$-$PvuII_a$ fragment (+3.5 to +2.1) labeled on the 3′ end of the PvuII site (Fig. 2B); 4) Primer extension of the DHFR-TS mRNA using a primer labeled on the 5′ end at the $PvuII_b$ site (Fig. 2B); 5) S1 nuclease mapping of the 3′ end of the 13 kb RNA using the 1.6 kb BglII-$SalI_b$ fragment (map position +.65 to −0.95) labeled on the 3′ end of the SalI site (Fig. 2B). Symbols: +, R1000 polyadenylated RNA plus S1 nuclease; −, S1 nuclease without RNA; P, undigested probe. DNA size markers (m) were MspI digested pBR322 DNA (not all fragment sizes are indicated). Arrows mark termini discussed in the text. Panel B. Restriction map and summary of results. The location of this segment within the R region is indicated by the coordinates above the map (see also Fig. 1). A, Sau3A; B, BglII, E, EcoRI, I, SphI; P, PstI; S, SalI; T, SstI; V, PvuII, X, XhoI. Restriction sites present more than once are distinguished by subscripts; for clarity only relevant Sau3A, SphI and PvuII sites are shown. S1 nuclease and primer extension products identified in panel A are shown below the map, with the star designating the labeled position and a vertical line the apparent map position of the observed product; 5′ or 3′ refer to the end that was mapped. Dashed S1 products indicate full length protection of the probe and thick lines denote predominant products. Vertical dashed lines linking S1 and primer extension products indicate associations discussed in the text. Panel C. Major RNAs are shown as wavy lines with open boxes on the 5′ ends designating miniexon sequences inferred from nuclease protection, primer extension and PCR amplification studies (Fig. 3). Sequences of polydinucleotide tracts are shown in the same sense as the DHFR-TS strand.

These data identify a 440 bp intertranscript region between the major 5′ downstream and 3′ DHFR-TS ends (Fig. 2C). This distance drops to 260 bp if minor ends are considered, and a small amount of RNA appears to span this region.

DHFR-TS subregion. Previous work has shown that the DHFR-TS mRNA originates and the upstream 13 kb RNA terminates within the 0.7 kb EcoRI-$SalI_b$ fragment of this region (Figs. 1 and 2B; 15), and that the DHFR-TS mRNA has two major 5′ ends mapping 253 and 388 bp to the right of the $PvuII_b$ site and two minor ends mapping 208 and 308 bp to the right of this site (Fig. 2B; 23). A small amount of fully protected probe was also observed (23).

Primer extension of the DHFR-TS mRNA produced two major products with apparent 5′ ends mapping 290 and 415 bp to the right of the $PvuII_a$ site (Fig. 2A), approximately 30−35 bp beyond the major S1 sites (Fig. 2B, marked by dashed lines). PCR amplification of DHFR-TS cDNA with the DHFR-TS primer and the miniexon-A primer produced two products whose sizes (280 and 400 bp) were similar to those obtained in primer extension (Fig. 3). Previous DNA sequencing of the predominant 280 bp product showed that it contains a complete miniexon on its 5′ end (19).

S1 nuclease mapping of the 3′ end of the 13 kb RNA (Fig. 2A) identified three major protected products whose termini map 125, 130 and 135 bp to the left of the $SalI_b$ site and two less abundant termini mapping 105 and 305 bp to the left of this site (Fig. 2B). These data identify a 445 bp intertranscript region between the most abundant 5′ DHFR-TS and 3′ 13 kb RNA ends (Fig. 2C). The intertranscript distance drops to 250 bp, if minor ends are considered, and a small amount of RNA may span this region (23).

Leftwards-transcribed RNAs of the divergent RNA subregion. The divergent RNA subregion (Fig. 1) contains the 5′ end of the 13 kb RNA and encodes four additional RNAs (1.7-*l*, 1.7-*r*, 6.2-*l* and 6.2-*r*) which are transcribed from opposite strands and extensively overlap one another (Fig. 1; 15). The suffixes *-l* (leftward) and *-r* (rightward) denote the direction of transcription, with the leftward RNAs being transcribed in the same direction as DHFR-TS.

Previous Northern blotting showed that the 5′ end of 13 kb RNA and 3′ end of the 1.7-*l* RNA map within the 0.8 kb $SalI_a$-$SphI_a$ fragment of this region (Fig. 4B; 15). S1 nuclease mapping of the 5′ end of the 13 kb RNA (Fig. 4A) revealed considerable heterogeneity, with the most abundant protected

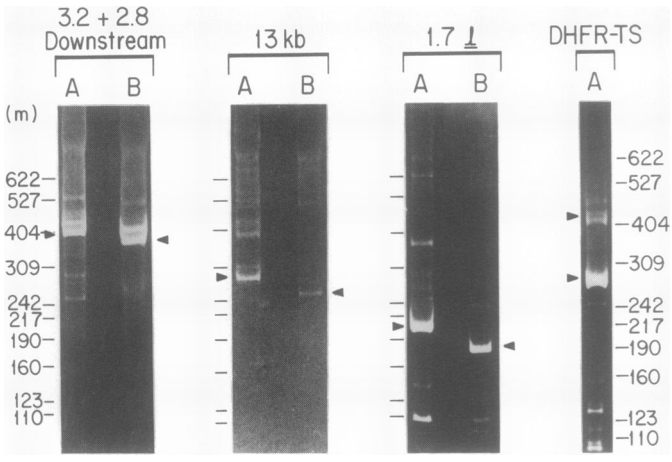


Figure 3. PCR amplification of R region RNAs. cDNAs generated by the indicated RNA-specific oligomers (described in the text and the legends to Figs. 2 and 4) were amplified using the RNA-specific oligomer in conjunction with miniexon primers A or B, corresponding to the 5′ and 3′ terminal portions of the *L. major* miniexon. Shown are ethidium bromide stained nondenaturing acrylamide gels of PCR products obtained for the downstream, 13 kb, 1.7-*l* and DHFR-TS RNAs. DNA size markers (m) were MspI digested pBR322 DNA (not all fragment sizes are indicated).

fragment constituting about 1/3 of the S1 products and terminating 260 bp to the right of the $SalI_{a'}$ site (Fig. 4B). Eight less abundant products mapped 370, 495, 510, 520, 540, 550, 610 and 630 bp to the right of the $SalI_{a'}$ site and a small amount of fully protected probe was observed. S1 mapping with a probe labeled instead at the $SstI_a$ site corroborated the existence and map positions of the multiple 5′ ends (data not shown).

Primer extension of the 13 kb RNA with a primer labeled at the $SalI_{a'}$ site (Fig. 4A, PE lane 1) produced a major product with an apparent 5′ end mapping 285 bp to the right of the $SalI_{a'}$ site, and minor products ending 320, 325, 360, 390, 510 and 530 bp to the right of this site (Fig. 4B). Primer extension with a second primer labeled at the 5′ 3nd 353 bp to the right of the $SalI_{a'}$ site (Fig. 4A, PE lane 2) was used to further investigate the multiple upstream termini observed for this RNA. Two major products with apparent 5′ ends mapping 533 and 548 bp to the right of the $SalI_{a'}$ site were observed, with additional ends mapping at 523, 628 and 638 bp.

PCR amplification of the 13 kb RNA with a primer whose 5′ was located at the $SalI_{a'}$ site in combination with miniexon primers A or B produced predominant products of 285 and 260 bp, respectively (Fig. 3), in addition to a heterogeneous series of larger products. PCR amplification with a primer whose 5′ end was located 353 bp to the right of the $SalI_{a'}$ site and miniexon primer A generated 5 products of 190, 200, 250, 275, 295 and 340 bp, whose size range was similar to that of primer extension products obtained with this primer (data not shown). The PCR amplification products hybridized with a probe specific for the 13 kb RNA and their occurrence was dependent on the presence of both RNA-specific and miniexon primers (data not shown). These results suggested that the predominant 5′ end of the 13 kb RNA bears the miniexon (Fig. 4B, dashed line), however due to the large number of S1, primer extension and PCR products correlations among the less abundant products were not obvious. The ability to amplify cDNAs mapping to the vicinity of the minor primer extension products suggests that many of these minor RNAs may contain miniexons, although sequencing will be required for confirmation.

S1 nuclease mapping of the 3′ end of the 1.7-*l* RNA (Fig. 4A) identified a predominant 3′ end mapping 515 bp to the left of the $PstI_a$ site, minor ends mapping 345 and 750 bp to the left of this site and a small amount of fully protected probe (Fig. 4B). These data identify a 660 bp region between the major termini of the 13 kb and 1.7-*l* RNAs (Fig. 4C). If minor ends are considered this distance decreases to 80 bp, and a small proportion of RNAs span this region.

S1 nuclease analysis previously indicated that the 5′ ends of the 1.7-*l* RNA map 150−155 bp to the right of the $SalI_c$ site of this region (Fig. 4B; 15). Primer extension of the 1.7-*l* RNA (Fig. 4A) produced a single product with an apparent 5′ end mapping 210 bp to the right of the $SalI_c$ site, 55−60 bp longer than the S1 products. PCR amplification of the 1.7-*l* cDNA with the primer extension oligonucleotide in combination with miniexon primers A or B produced a major product of 205 bp with minexon primer A and 175 bp with miniexon primer B (Fig. 3). These products hybridized with a probe specific for this region and their occurrence was dependent on having both RNA- and miniexon-specific primers in the reaction (data not shown). These data suggest that the 1.7-*l* RNA has a miniexon on its 5′ end.

Previous Northern blot analysis mapped the 3′ end of the 6.2-*l* RNA to the 0.9 kb $SalI_c$-$PstI_b$ fragment of this region (Fig. 4B; 15). S1 nuclease mapping of the 3′ end of the 6.2-*l* RNA (Fig. 4A) identified a major 3′ end terminating 120 bp to the left of the $SphI_c$ site and minor termini mapping 165 and 280 bp to the left of this site (Fig. 4B). These analyses identified a 470 bp intertranscript region between the major 5′ 1.7-*l* and 3′ 6.2-*l* RNA termini, although if minor ends are considered this distance is 310 bp (Fig. 4C).

Rightwards RNAs of the divergent RNA region. Previous Northern blot analysis mapped the 5′ end of the 1.7-*r* RNA to the 0.55 kb $SphI_a$-$SphI_b$ fragment of this region (Fig. 4D; 15). Multiple attempts to map the 5′ end of the 1.7-*r* RNA by S1 nuclease protection with the 1.4 kb $SstI_a$-$SalI_b$ or 1.1 kb $SstI_a$-$PstI_a$ fragments labeled on their 5′ end at the $SalI_b$ or $PstI_a$ sites (Fig. 4B) were unsuccessful, even though mixing experiments with a downstream RNA probe and the 1.7-*r* RNA $SstI_a$-$PstI_a$ probe yielded a protected fragment corresponding to the major downstream RNA 5′ end (Fig. 4A, lanes labeled Mixing). These results were perplexing, since S1 nuclease mapping of the 1.7-*r* RNA with a probe spanning the internal portion of this RNA ($PstI_a$-HindIII, Fig. 4B) and overlapping the $SstI_a$-$SalI_b$ probe successfully yielded full length protection of the *Leishmania* sequences (Fig. 4D), as would be predicted for a contiguous 1.7-*r* RNA spanning this segment. This experiment also suggests that the formation of RNA-RNA duplexes cannot account for the failure to protect a fragment located at the 5′ terminus. DNA sequence analysis of this region (below) revealed no unusual sequence organization or elements that could account for these discrepant S1 results. Previously, a probe labeled at the $SalI_c$ site identified 3′ termini mapping 620−630 bp to the right of the $SalI_c$ site (Fig. 4D; 15).

In contrast to the results obtained by S1 nuclease protection, primer extension mapping of the 5′ end of the 1.7-*r* RNA with a primer labeled at the $PstI_a$ site (Fig. 4B) yielded three comparably abundant products (Fig. 4D) with apparent 5′ ends mapping 120, 135 and 140 bp to the left of the $PstI_a$ site. A second primer situated 95 bp to the right of the $PstI_a$ site

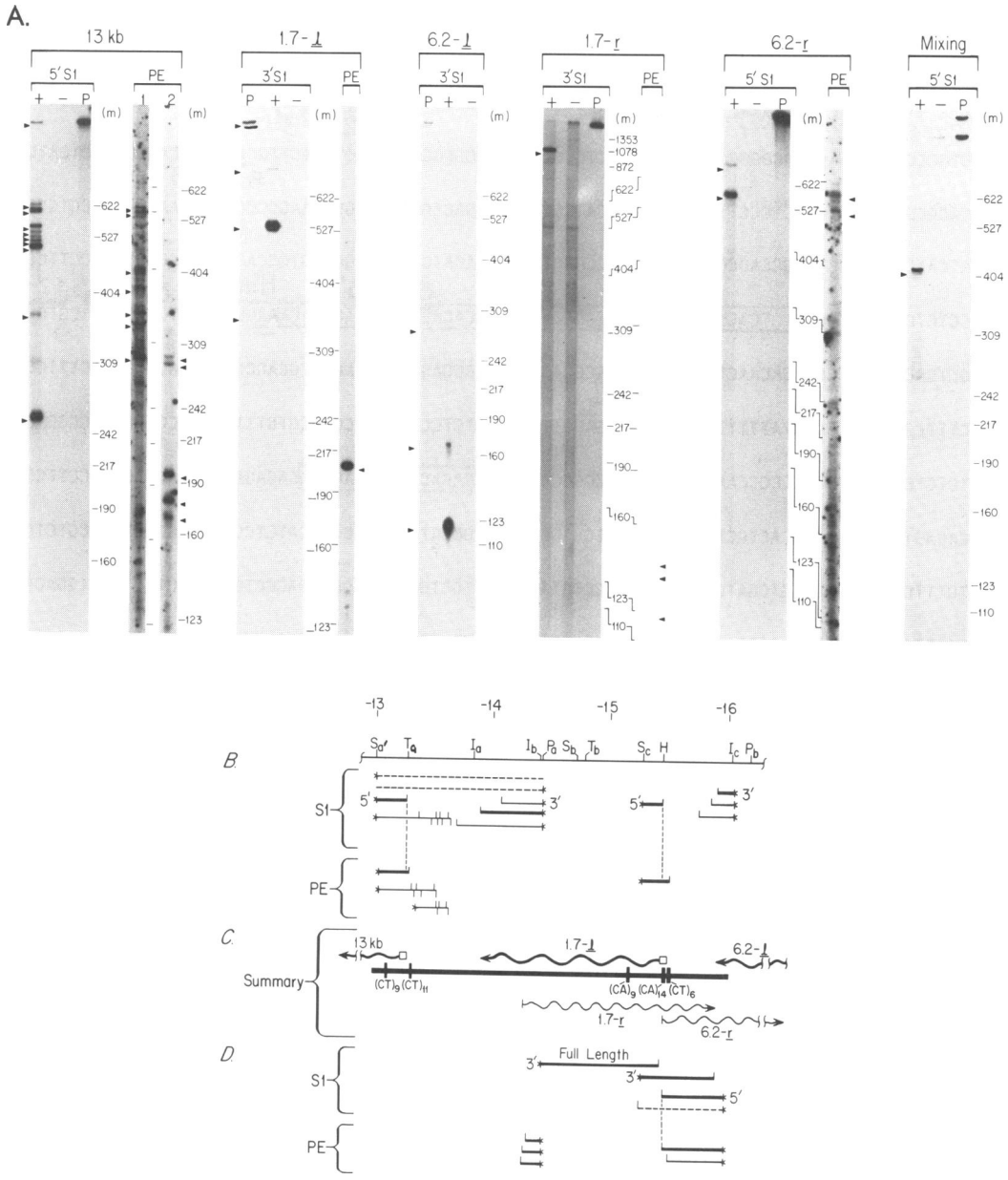
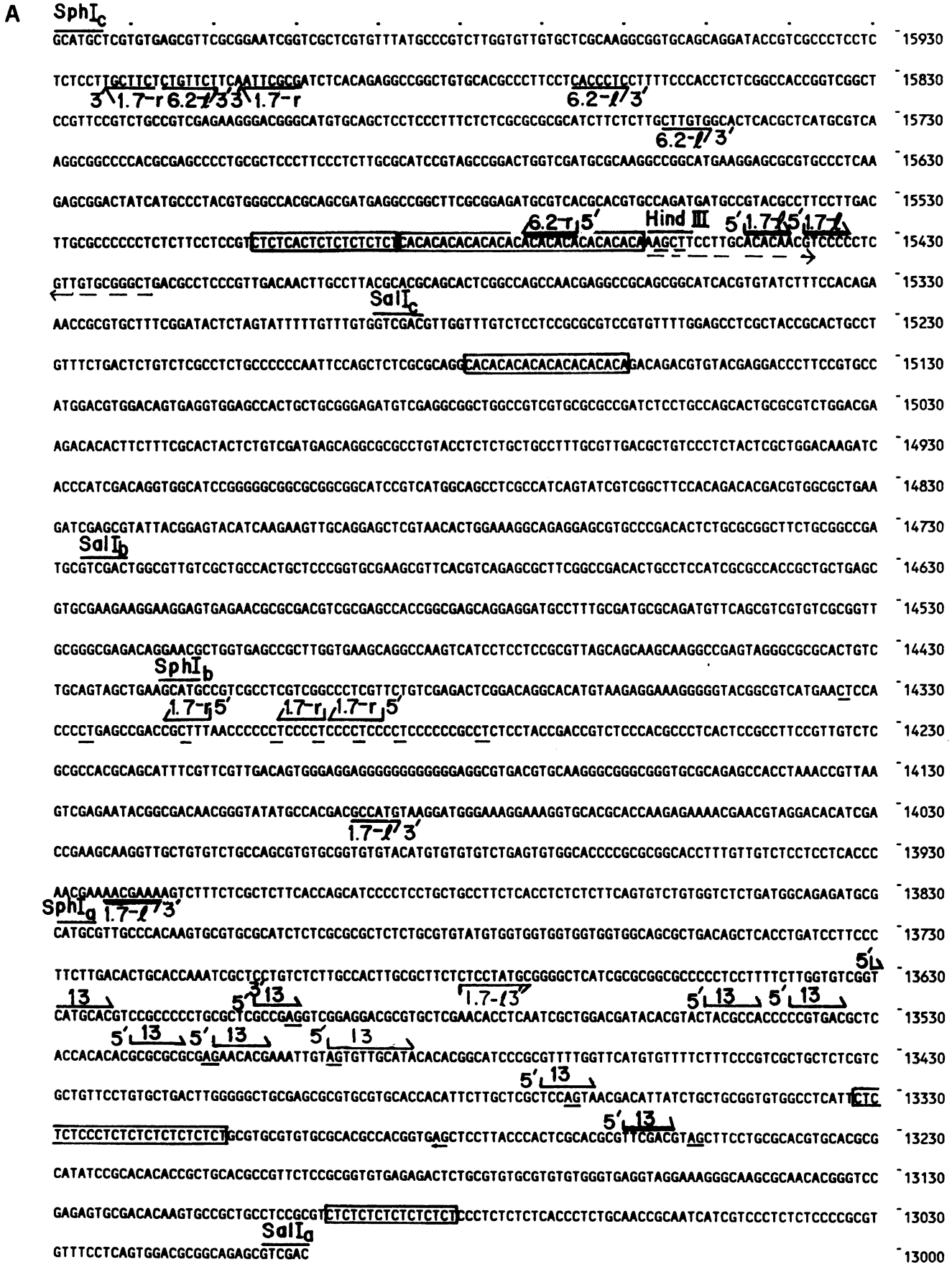


Figure 4. S1 nuclease and primer extension analysis of RNA termini within the divergent RNA subregion. Panel A. S1 nuclease and primer extension analysis. For locations of sites refer to the map shown in panel B. Symbols: +, 2 μg R1000-3 poly(A)$^+$ RNA plus S1 nuclease; −, S1 nuclease alone; P, undigested probe. 5'S1, 3'S1, and PE refer to S1 nuclease mapping of 5' and 3' termini of RNA and primer extension, respectively. DNA size markers (m): MspI digested pBR322 DNA (not all fragment sizes are indicated). Additional higher molecular weight markers (ϕX HaeIII fragments) are indicated where appropriate. 13 kb: left panel, S1 nuclease mapping of the 5' end of the 13 kb RNA (lanes labeled 13 kb, 5'S1) with the 1.4 kb $SalI_a$-$PstI_a$ fragment (positions −13.0 to -14.4) labeled on the 5' end of the SalI site; right panel, primer extension of the 13 kb RNA with primers labeled at the Sal_a site (lane labeled 13 kb, PE 1) or at the position 353 nt to the right of the $SalI_a$ site (lane labeled 13 kb, PE 2); 1.7-*l*: left panel, S1 nuclease mapping of the 3' end of the 1.7-*l* RNA (lanes labeled 1.7-*l*-3' S1) with the 1.1 kb $PstI_a$-$SstI_a$ fragment (position −13.3 to −14.4) labeled on the 3' end of the PstI site; right panel, primer extension of the 1.7-*l* RNA with a primer labeled on the 5' end at the $SalI_c$ site (lane labeled 1.7-*l*, PE); 6.2-*l*: S1 nuclease mapping of the 3' end of the 6.2-l RNA with a 3.75 kb $SphI_c$-BamHI fragment containing 0.75 kb of *Leishmania* DNA ($SalI_c$-$SphI_c$, positions −15.2 to −16.0) and 3.0 kb of vector DNA labeled on the 3' end of the SphI site (lane labeled 6.2-*l*, S1); 1.7-*r*: left panel: S1 nuclease mapping of an internal segment of the 1.7-*r* RNA (lanes labeled 1.7-*r*, 3'S1) with a 5.0 kb $PstI_a$-BamHI fragment containing 1.0 kb of *Leishmania* DNA ($PstI_a$-HindIII, positions −14.4 to −15.4) and 4.0 kb of vector DNA, labeled at the 3' end of the PstI site; right panel: primer extension of the 1.7-*r* RNA with a primer labeled on the 5' end at the $PstI_a$ site (lane labeled 1.7-*r*, PE); 6.2-*r*: left panel, S1 nuclease mapping of the 5' end of the 6.2-*r* RNA with a 3.75 kb $SphI_c$-BamHI fragment containing 0.75 kb of *Leishmania* DNA ($SalI_c$-$SphI_c$, positions −15.2 to −16.0) and 3.0 kb of vector DNA labeled on the 5' end of the $SphI_c$ site (lane labeled 6.2-*r*, 5' S1); right panel, primer extension of the 6.2-*r* RNA with a primer labeled on the 5' end at the $SphI_c$ site (lane labeled 6.2-*r*, PE); Mixing: Control S1 mapping experiment (lanes labeled mixing): S1 probes designed for the 5' end of the 1.7-*r* kb RNA probes (1.15 kb $SstI_a$-$PstI_a$ fragment (positions −13.2 to −14.4) labeled on the 5' end of the PstI site) and 5' end of the downstream RNAs (see Fig. 2B) were mixed and used for standard S1 nuclease digestion. The pattern shown is the same as that obtained with the downstream probe alone (see Fig. 2A). Panel B. Restriction map and summary of results obtained for leftwards RNAs. Results are shown using the conventions described in Fig. 2B. H, HindIII; I, SphI; P, PstI; S, SalI; T, SstI. Sites present more than once are distinguished with subscripts. S1 products of the 5' end of the 1.7-*l* RNA were previously determined (15). 5' and 3' refer to the ends that were mapped. Panel C. Summary of Transcript Organization. The RNAs of this portion of the divergent region are shown using the conventions described in Fig. 2C. Only the major RNAs are shown. Vertical tick marks represent termini. Panel D. Summary of results obtained for rightwards RNAs. Data are shown using the conventions described in Fig. 2B. S1 products of the 3' end of the 1.7-*r* RNA were previously determined (15).


A
SphI c
SalI c
SalI b
SphI b
SphI a
SalI a
Hind III
6.2-r
6.2-l
1.7-r
1.7-l
1.7-13
13
5'
3'
15930
15830
15730
15630
15530
15430
15330
15230
15130
15030
14930
14830
14730
14630
14530
14430
14330
14230
14130
14030
13930
13830
13730
13630
13530
13430
13330
13230
13130
13030
13000

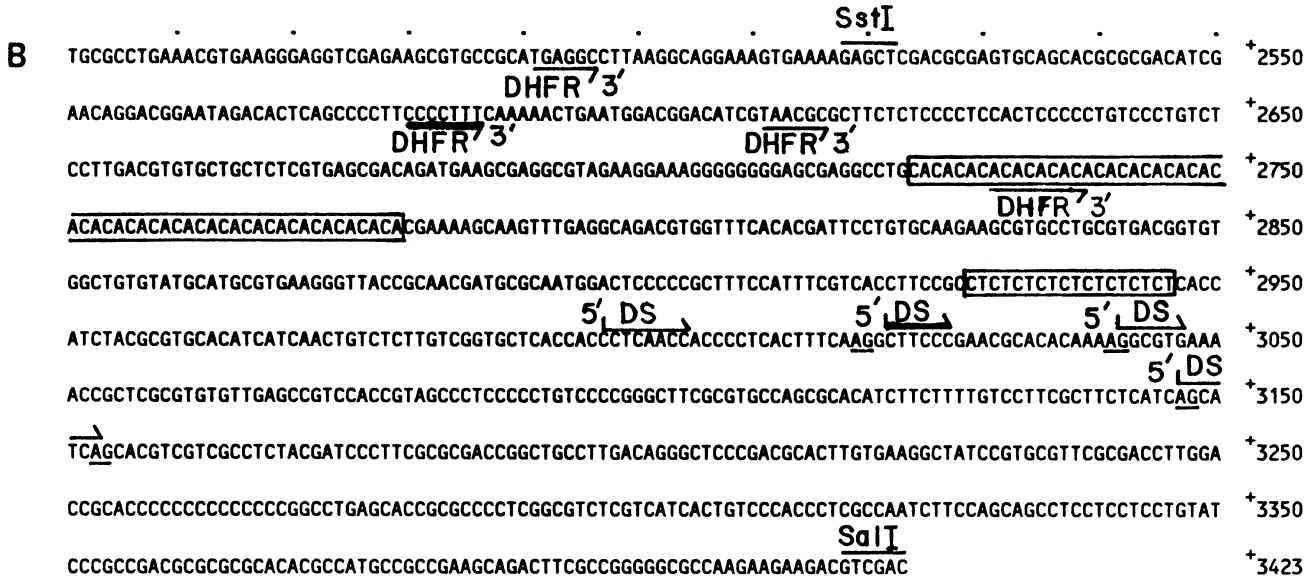


Figure 5. DNA sequence encompassing intertranscript regions of the R region. Panel A. Sequence of the 3031 bp $SalI_a$'-$SphI_c$ segment of the divergent RNA subregion (Figs. 1, 4B). The same strand as DHFR-TS is shown (the sequence shown is in reverse orientation to map shown in Fig. 4B). The $SalI_a$' site was mapped at position −13 kb (Fig. 4B). RNA termini are shown with 5' ends displayed above the sequence and 3' ends below, with arrows denoting the direction of transcription. Abundant termini are indicated by heavy arrows. All termini shown were obtained by S1 mapping, with the exception of the 1.7-*r* RNA 5' end which was determined solely by primer extension as discussed in the text. Potential 3' splice acceptor AG sites are underlined (CT for RNAs from the opposite strand). Polydinucleotide tracts are boxed. A potential hairpin structure is shown by a dashed line above the sequence near the HindIII site. Panel B. Sequence of the region between the DHFR-TS and downstream RNAs (Fig. 2B), nucleotides 2451−3423. Termini are shown as described in panel A. The sequence is shown in opposite orientation to the map shown in Fig. 2B.

produced three major primer extension products mapping to these same apparent positions (data not shown). The distance between the 5' and 3' ends of the 1.7-*r* RNA, mapped by primer extension and S1 nuclease, respectively, is approximately 1.6 kb, consistent with the size estimated by Northern blot analysis.

Previous Northern blots mapped the 5' end of the 6.2-*r* RNA to the 0.9 kb $SalI_c$-$PstI_b$ fragment of this region (Fig. 4B; 15). S1 nuclease mapping of the 5' end of the 6.2-*r* RNA (Fig. 4A) mapped a major 5' end 560 bp to the left of the $SphI_c$ site (Fig. 4D), as well as a small amount of full length protection of the *Leishmania* sequences within this probe (750 nt fragment, Fig. 4A). Unlike the leftward transcribed RNAs which contain gaps between the adjacent major transcripts, the regions encoding the 1.7-*r* and 6.2-*r* RNAs overlapped by 450 bp (Fig. 4C). Primer extension of the 6.2-*r* RNA identified two products (Fig. 4A) with apparent 5' ends mapping 500 and 550 bp to the left of the $SphI_c$ site (Fig. 4D). The similarity in size of the major 6.2-*r* S1 and primer extension products (S1: 560 nt, PE: 550 nt) suggested that this RNA lacks a miniexon.

PCR amplification was attempted to determine if miniexon sequences are present on the 5' ends of the 1.7-*r* and 6.2-*r* RNAs. Despite varying the reaction conditions and testing two primers specific for each RNA, a large number of products were observed whose sizes did not vary appropriately with the miniexon A and B primers (data not shown). Thus, current data do not reveal the presence of miniexon sequences on the rightwards RNAs arising from the divergent region.

DNA sequence analysis

The DNA sequences of the regions encompassing the termini mapped above were obtained (Figs. 5A, 5B; 23) and analyzed for shared or conserved sequences which could mark functionally important signals. All major 5' termini identified by S1 nuclease analysis mapped close to AG dinucleotides required for trans-splicing (Fig. 5). Most minor termini revealed by S1 protection analysis also mapped in close proximity to AG residues, although exceptions were noted (13K 5' ends located at positions −13447, −13540, −13550; downstream 5' end located at position +2900; Fig. 5).

Table 1 shows an alignment of the four miniexon addition sites of the leftwards RNAs and three other *Leishmania* genes. Most *Leishmania* splice acceptor sites are flanked by pyrimidine-rich regions on both the 5' and 3' sides, whereas generally eukaryotic splice acceptor sequences show pyrimidine-richness only on the 5' side (33). This may reflect the fact that the splice acceptor sequence in trypanosomatids is invariably non-coding. Sequences related to the hexanucleotide CTTCC(T/C) were found immediately 3' of the AG marking the splice acceptor site for all seven transcripts shown in Table 1.

One common sequence motif identified in the vicinity of the AG splice acceptor dinucleotide was the pentamer CGCAC. This or related sequences were found in more than 20 positions, both 5' and 3' of the AG dinucleotide. This sequence is complementary to the precursor RNA which donates the miniexon (termed the med-RNA) near the 5' splice donor site (Table 1). Layden and Eisen (34) have proposed that base-pairing interactions may serve to guide the site of the trans-splicing reaction, although the regions of complementarity that they identified were variable and frequently adjacent to the region complementary to CGCAC. However, the CGCAC sequence was found both upstream and downstream of the splice acceptor site as well as throughout the sequenced regions of *Leishmania* DNA, and other additional

Table 1. Sequence alignments of the 3'splice junction of *Leishmania* mRNAs

```
DHFR-TS GCTTGACGCATACGCGACGAATTCGAAAG/CTCACCTCATTCCTCCCTCCTCACACCATCA
Dwnstrm ACCACCCTCAACCACCCCTCACTTTCAAG/GCTTCCCGAACGCACACAAAAGGCGTGAAAA
1.7-l   CTTCCTCCGTCTCTCA(CT)6(CA)15AG/CTTCCTTGCACACAACGTCCCCCTCGTTGTG
13K     CCTTACCCACTCGCACGCGTTCGACGTAG/CTTCCTGCGCACGTGCACGCGCATATCCGCA

hsp70   CCCCCCCCTATCCACCAAACCACACGCAG/GATCCTAAACACGCACTCGCACTCAAGCTGT
α-TUB   CTCTCCCCTCTCCCCGCACACACACGCAG/TCCTTCGCTTTCACTCTTCGAACAAACACCT
ß-TUB   GTCCATACCACCGGCCACCCCCTACCCAG/TATTCGTCCCCCAGCACTCTTCACTACACAT

L. enrietti med-RNA:  TATTG/GTATGCGAA
                          ||||
                        CACGC-5'
```

The splice junctions associated with the major termini of the RNAs listed are shown, with a '/' marking the splice site, pyrimidine rich regions marked by underlining, and the CGCAC motifs discussed in the text marked in bold type. Sequences are DHFR-TS, Dwnstrm (3.2 and 2.8 kb RNAs, Fig. 2B), 13K and 1.7-*l* (Fig. 4C), all from the R region of *L. major* (23; this work); HSP70, 70K heat shock protein homolog of *L. major* (53); α-TUB and β-TUB, α- and β-tubulins respectively of *L. enriettii* (42); med-RNA, the miniexon precursor RNA sequence of *L. enriettii* (54). The '/' in this sequence denotes the splice donor site. The sequence shown of the med-RNA is identical in all *Leishmania* species studied thus far (55).

sequences potentially capable of base-pairing with the med-RNA were observed (not shown). These considerations suggest that the assignment of a role in trans-splicing to these sequences will require functional tests.

Analysis of the DNA flanking the 3' ends of the R region RNAs did not identify sequences which are conserved amongst these regions. These regions did not exhibit similarity with the prototypic polyadenylation signals of higher eukaryotes or yeast (35–37). The 3' end regions also tended to be flanked by pyrimidine rich segments, and short direct repeats specific to a given transcript were commonly found in the vicinity of the 3' ends.

Curiously, the predominant 3' ends of the 1.7-*r* and 6.2-*l* RNAs localized to similar positions, as did the 5' ends of the 1.7-*l* and 6.2-*r* RNAs. Using the method of Zucker (38), we examined these two regions for potential secondary structures which might exist in DNA or RNA. A potential hairpin structure with an estimated $\hat{-}$G of −19.7 kCal was observed only for the DNA surrounding the divergent 5' ends.

The intertranscript regions were searched for conserved or consensus sequence elements found in other species, however no similarities were identified which displayed a conserved position relative to the 5' or 3' RNA termini. Specific searches for elements such as TATAAT and CAAT boxes (39), binding sites for the transcription factor SP1 (GGGGCGG; 40), and the yeast splice acceptor branch site TACTAAC (41), among others, were unsuccessful other than a potential SP1 binding site (GGGGCTG) 150 bp upstream of the major 5' end of the 13 kb RNA.

Comparison of the intertranscript regions with RNA-coding regions revealed polydinucleotide tracts (>10 nt) within or in close proximity to the intertranscript regions (Figs. 2C, 4C, 5A and 5B). Poly-(CA) tracts were found between the downstream and DHFR-TS RNAs ($(CA)_{28}$) and immediately upstream of the smaller major DHFR-TS 5' end ($(CA)_{20}$); this tract falls within the larger DHFR-TS 5' end; Figs. 2B and 5A). Poly-(CA) tracts were also found in the intergenic regions of the tandemly repeated tubulin genes of *L. enriettii* (42). Curiously, poly-(CA) is similar to the med-RNA complementary sequence CGCAC discussed above.

Protein-coding potential of the 1.7 kb RNAs

The region encompassing the overlapping opposite-strand 1.7-*l* and 1.7-*r* RNAs was analyzed for open reading frames (ORFs), assuming initiation only at AUGs and the standard genetic code. The largest ORFs within the 1.7-*l* RNA were 210, 252 and 261 bp in length, with the most 5' of these beginning 900 nucleotides from the 5' end of the RNA and preceded by at least 6 AUGs and 4 smaller ORFs. Quantitative analysis of codon usage bias of these ORFs revealed them to be compositionally random and not in agreement with that of other *Leishmania* proteins (20; unpublished). The 1.7-*r* RNA exhibited one large 579 nt ORF, however this ORF began 650 bases from the 5' end of the mRNA, was proceeded by 10 AUGs and 5 smaller ORFs, and the predicted protein did not conform to the expected *Leishmania* protein codon bias. These results suggested that it was unlikely that either of the two 1.7 kb RNAs encoded proteins.

DISCUSSION

To further our understanding of the transcriptional organization of the DHFR-TS gene and surrounding R region, we have mapped the termini of 8 RNAs transcribed from the amplified R region of *Leishmania major* and sequenced the DNA encompassing 6 inter-transcript regions and two opposite strand transcripts. Heterogeneity at both 5' and 3' ends of the RNAs was commonly observed, the most extreme case being the 13 kb RNA which utilized at least eight 5' and four 3' termini. Whether the 5' and 3' ends are determined independently, as for the mouse DHFR gene (43), or in limited combinations has not been addressed. Our data show that the predominant termini of the leftwards RNAs contain miniexons at their 5' ends. Many of the minor termini appear to contain miniexons as well, and thus must be the result of alternative trans-splicing. Although alternative cis-splicing can play important functional roles in the expression of higher eukaryotic genes, this does not appear to be the case for the DHFR-TS mRNA since the 5' heterogeneity is invariant in promastigotes and amastigotes (23).

While many of the 5' ends of the R region RNAs contain miniexons, it is clear that some do not. Two of these are major RNAs, the rightwards-transcribed 1.7-*r* and 6.2-*r* RNAs, which partially overlap. Since one potential role of the miniexon is to provide the 5' cap structure required for translation, it would be expected that the RNAs lacking miniexons would be non-coding, and this appears to be the case for the 1.7-*r* RNA. However, the DNA sequence of the opposite-strand 1.7-*l* RNA, which bears the miniexon, also did not reveal clear evidence for an encoded protein. Both of these RNAs have previously been shown to occur on polysomes (15). Several mechanisms could account for these results: it is possible that certain termination codons are being suppressed, or that mechanisms such as nuclear RNA editing or RNA modification mediated by pairing of the antisense transcripts (44,45) could transform the genomic sequence into a protein-coding readiang frame. If present, these modifications were not detected by S1 mapping using genomic DNA probes and have not been previously reported in trypanosomatids. In support of the sequencing results, two-dimensional gel analysis of proteins of R region-amplified *Leishmania* failed to reveal any over-expressed proteins other

than DHFR-TS (46). Noncoding polysomal RNAs in trypanosomes have been reported (14), though precedents for non-translated polysomal RNAs are rare in other species (47–49).

In keeping with work from other trypanosome genes, other than limited similarity in the vicinity of the miniexon splice acceptor sites no conservation of universal sequence 'motifs' at key positions around transcripts of the DHFR-TS region was observed. This is not surprising since one common eukaryotic signal, that marking poly-adenylation, has been shown to be non-functional in transfected derivatives of the R region (19). One feature of the intertranscript regions was the presence of polydinucleotide repeats, most commonly poly-(CA), which are also observed in other *Leishmania* genes. Poly-(CA) tracts, which are capable of adopting a lefthanded DNA conformation, can enhance transcription in mammalian cells and when transcribed form cytoplasmic 'Z-RNA' in trypanosomatids (50; 51; 52). It is possible that the dinucleotide tracts constitute loose signals for transcription or processing.

Mapping of the transcripts adjacent to the DHFR-TS mRNA has allowed us to consider these RNAs in the context of the current 'polycistronic' model for transcription in trypanosomatids. Transcripts of the R region exhibit several hallmarks associated with this model: 1) A series of closely juxtaposed RNAs transcribed from the same strand (6.2-*l*, 1.7-*l*, 13K, DHFR-TS, downstream RNAs), all bearing the trans-spliced miniexon. 2) Occasional protection of inter-transcript regions in S1 nuclease protection assays of total cellular poly-adenylated RNA preparations, not enriched for transcriptional intermediates. Had nuclear RNA been employed, the occurrence and amount of RNA protection observed might have been greater. 3) The occurrence of apparently non-functional RNAs within this putative 'polycistronic' domain, minimally the 1.7-*l* and 1.7-*r* RNAs and perhaps many of the others. At a structural level, these data are consistent with evidence obtained from other trypanosomatid loci thought to exhibit polycistronic transcription by these and additional criteria. In contrast, the transcripts proceeding in the rightwards direction lack many of these features, and may arise from an independent mechanism.

Although consistent with the polycistronic transcriptional model, our data do not rule out other models for transcription of the R region. Independent transcription of each RNA could occur, assuming that initiation was sufficiently far upstream to accommodate the trans-splicing mechanism as well as yielding at least some RNAs which span the inter-transcript regions. The emerging methods of DNA transfection of parasites promise to resolve this issue, as a polycistronic DHFR-TS transcription model predicts that transcription should be dependent upon specific promoters located at some considerable distance upstream of the mature DHFR-TS mRNA.

ACKNOWLEDGEMENTS

We thank Tom Ellenberger for discussions and unpublished data, and Deborah Dobson, Jon LeBowitz and Kim Nelson for reading this manuscript. Supported by Public Health Service grant AI-2903 from the National Institutes of Health to SMB and BRSG S)7 RR 05381-27 to Harvard Medical School. GMK and KZ were supported in part by training grants 5T-32-GM07306 and 5T-32-GM07196 through much of this work. SMB is a Burroughs-Wellcome Scholar in Molecular Parasitology.

REFERENCES

1. Sutton, R.E. and Boothroyd, J.C. (1986) Cell **47,** 527–535.
2. Murphy, W.J., Watkins, K.P. and Agabian, N. (1986) Cell **47,** 517–525.
3. DeLange, T., Michels, P.A.M., Veerman, H.J.G., Cornellisen, A.W.C.A. and Borst, P. (1984) Nucleic Acids Res. **12,** 3777–3790.
4. Walder, J.A., Eder, P.S., Engman, D.M., Brentano, S.T., Walder, R.W., Knutzon, D.S., Dorfman, D.M. and Donelson, J.E. (1984) Science **233,** 569–571.
5. Donelson, J.E. (1989) In Berg, D.E. and Howe, M.M. (eds.), Mobile DNA, Am. Soc. Microb., Washington, D.C., pp.763–781.
6. Simpson, L. and Shaw, J. (1989) Cell **57,** 355–366.
7. Borst, P. (1986) Ann. Rev. Biochem. **55,** 701–732.
8. Clayton, C.E. (1988) Genetic Engineering **7,** 1–56.
9. Gonzalez, A., Lerner, T.J., Huecas, M., Sosa-Pineda, B., Nogueira, N. and Lizardi, P.M. (1985) Nucleic Acids Res. **13,** 5789–5804.
10. Muhich, M.J. and Boothroyd, J.C. (1988) Mol. Cell. Biol. **8,** 3837–3846.
11. Johnson, P.J., Kooter, J.M. and Borst, P. (1987) Cell **51,** 273–281.
12. Tschudi, C. and Ullu, E. (1988) EMBO J. **7,** 455–463.
13. Pays, E., Tebabi, P., Pays, A., Coquelet, H., Revelard, P., Salmon, D. and Steinert, M. (1989) Cell **57,** 835–845.
14. Aline, R.F., Jr., Scholler, J.K. and Stuart, K. (1989) Mol. Bioch. Parasit. **32,** 169–178.
15. Kapler, G.M. and Beverley, S.M. (1989) Molec. Cell. Biol. **9,** 3959–3972.
16. Bellofatto, V. and Cross, G.A.M. (1989) Science **244,** 1167–1169.
17. Laban, A. and Wirth, D.F. (1989) Proc. Natl. Acad. Sci. USA **86E,** 9119–9123.
18. Laban, A., Tobin, J.F., de Lafaille, M.A.C. and Wirth, D.F. (1990) Nature **343,** 572–574.
19. Kapler, G.M., Coburn, C.M. and Beverley, S.M. (1990) Mol. Cell. Biol. **10,** 1084–1094.
20. Beverley, S.M., Ellenberger, T.E. and Cordingley, J.S. (1986) Proc. Natl. Acad. Sci. USA **83,** 2584–2588.
21. Grumont, R., Washtien, W.L., Caput, D. and Santi, D.V. (1986) Proc. Natl. Acad. Sci. USA **83,** 5387–5391.
22. Beverley, S.M., Coderre, J.A., Santi, D.V. and Schimke, R.T. (1984) Cell **38,** 431–439.
23. Kapler, G.M., Zhang, K. and Beverley, S.M. (1987) Nucleic Acids Res. **15,** 3369–3383.
24. Mitchell, P.J., Carothers, A.M., Han, J.H., Harding, J.D., Kas, E., Venolia, L. and Chasin, L.A. (1986) Mol. Cell. Biol. **6,** 425–440.
25. Masters, J.N. and Attardi, G. (1985) Mol. Cell. Biol. **5,** 493–500.
26. Gasser, C.S. and Schimke, R.T. (1986) J. Biol. Chem. **261,** 6938–6946.
27. Dynan, W.S., Sazer, S., Tjian, R. and Schimke, R.T. (1986) Nature **319,** 246–248.
28. Coderre, J.A., Beverley, S.M., Schimke, R.T. and Santi, D.V. (1983) Proc. Natl. Acad. Sci. USA **80,** 2132–2136.
29. Cathala, G., Savouret, J.F., Mendez, B., West, B., Karin, M., Martial, J.A. and Baxter, J.D. (1983) DNA **2,** 329–335.
30. Aviv., H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA **69,** 1408–1412.
31. Beverley, S.M. (1990), in Ausubel, F., Brent, R., Kingston, R., Moore, D., Seidman, J., Smith, J.A., and Struhl, K., *Current Protocols in Molecular Biology*, Greene Publishing Associates, New York., pages 15.4.1–15.4.6.
32. Messing, J. (1983) In Wu, R., Grossman, L. and Moldave, K. (eds.), Methods Enzymology, Academic Press, New York, Vol. 101, pp.20–78.
33. Mount, S. (1982) Nucleic Acids Res. **10,** 459–472.
34. Layden, R.E. and Eisen, H. (1988) Mol. Cell. Biol. **8,** 1352–1360.
35. Birnsteil, M.L., Busslinger, M. and Strub, K. (1985) Cell **41,** 349–359.
36. Zaret, K.S. and Sherman, F. (1982) Cell **28,** 563–573.
37. Zaret, K.S. and Sherman, F. (1984) J. Mol. Biol. **176,** 107–135.
38. Jacobson, A., Good, L., Simonetti, J. and Zucker, M. (1984) Nucleic Acids Res. **12,** 45–52.
39. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. **50,** 349–383.
40. Dynan, W.S. and Tjian, R. (1983) Cell **35,** 79–87.
41. Langford, C.J. and Gallwitz, D. (1983) Cell **33,** 519–527.
42. Landfear, S.M., Miller, S.I. and Wirth, D.F. (1986) Mol. Bioch. Parasit. **21,** 235–245.
43. Yen, J.-Y.J. and Kellems, R.E. (1987) Mol. Cell. Biol. **7,** 3732–3739.
44. Kimelman, D. and Kirschner, M.W. (1989) Cell **59,** 687–696.
45. Bass, B.L. and Weintraub, H. (1989) Cell **55,** 1089–1098.
46. Ellenberger, T.E. (1989) Ph.D. Thesis **1,** 1–196.
47. Gross, M.K. and Merrill, G.F. (1989) Proc. Natl. Acad. Sci. USA **86,** 4987–4991.
48. Ballinger, D.G. and Pardue, M.L. (1983) Cell **33,** 103–114.
49. Thomas, G.P. and Mathews, M.B. (1984) Mol. Cell. Biol. **4,** 1063–1072.
50. Johnston, B.H. and Rich, A. (1986) Cell **47,** 713–724.

51. Zarling, D.A., Calhoun, C.J., Hardin, C.C. and Zarling, A.H. (1987) Proc. Natl. Acad. Sci. USA **84,** 6117–6121.
52. Hamada, H., Seidman, M., Howard, B.H. and Gorman, C.N. (1984) Mol. Cell. Biol. **4,** 2622–2630.
53. Lee, C.C., Wu, X., Gibbs, R.A., Cook, R.G., Muzny, D.M. and Caskey, C.T. (1988) Science **239,** 1288–1291.
54. Miller, S.I., Landfear, S.M. and Wirth, D.F. (1986) Nucleic Acids Res. **14,** 7341–7360.
55. Scholler, J.K., McArdle, S., Reed, S. and Kanemoto, R. (1989) Nucleic Acids Res. **17,** 7999–7999.