© 1993 Oxford University Press

Nucleic Acids Research, 1993, Vol. 21, No. 3 687–694

Structural organization of the maxicircle variable region of *Trypanosoma brucei:* identification of potential replication origins and topoisomerase II binding sites

Peter J.Myler, David Glick+, Jean E.Feagin, Tony H.Morales and Kenneth D.Stuart*
Seattle Biomedical Research Institute, 4 Nickerson Street, Seattle, WA 98109-1651, USA

Received September 9, 1992; Revised and Accepted December 18, 1992 EMBL accession no. Z15118

ABSTRACT

The maxicircle of the parasitic protozoan *Trypanosoma brucei*, one component of the mitochondrial genome, has size differences among isolates that localize to the variable region (VR) between the ND5 and 12S rRNA genes. We present here the nucleotide sequence of this entire region, thus completing the sequence of the maxicircle genome. We also find heterogeneously sized transcripts from throughout most of the VR. The VR has three distinct sections, each with characteristic repeated sequences. The repeated sequences in two sections are short and highly reiterated; the intra-species size variation occurs within this region. The third section contains non-repetitive sequences and a large duplication immediately upstream of the 12S rRNA gene. Two repeat units within section I contain a sequence that has homology to the DNA replication origin of minicircles. This region also contains sequences with homology to topoisomerase II binding and cleavage sites. These findings suggest a role for the VR in DNA replication of the maxicircle.

INTRODUCTION

The mitochondrial genome of trypanosomatids is composed of a complex network containing two types of concatenated circular molecules (for reviews see 1,2). In *Trypanosoma brucei* there are 5,000–10,000 copies of 1 kb minicircles which fall into ~ 400 different sequence classes, and 50 copies of the 21–27 kb maxicircle. The maxicircle encodes a number of the genes which are normally found in the mitochondrial genomes of other organisms, in addition to other components of the electron transport chain, which are often nuclearly encoded. These include the 12S and 9S rRNAs, apocytochrome *b*, cytochrome *c* oxidase subunits I, II, and III, NADH dehydrogenase subunits 1, 4, 5, 7, 8 and 9, ATPase 6, ribosomal protein S12, and several unidentified genes (MURF1, MURF2, CR3, CR4, CR5 and US). Interestingly, in contrast to other organisms, mitochondrial tRNAs in *T. brucei* are nuclearly encoded (3). While the location of trypanosomatid mitochondrial genes is fairly conventional, their transcription and processing is not. Transcription is polycistronic (4) and the abundance of several transcripts is differentially regulated between life-cycle stages (5–7), and many of the transcripts are extensively edited by post-transcriptional addition and/or deletion of uridines (see 8,9 for reviews). The information to direct RNA editing is largely encoded on the minicircles in the form of guide (g)RNAs (10).

Although the maxicircles from different trypanosomatids encode the same genes, their size varies, both between species and among different stocks of the same species. Some of the inter-species variation is due to differences in the genomic sequence for the edited genes, with *T. brucei* containing fewer thymidines than *Leishmania tarentolae* and *Crithidia fasciculata*. However, the major area of divergence between species, and essentially all the divergence within species, occurs in the region between the ND5 and 12S rRNA genes. This region, known as the variable (or divergent) region (VR) varies in length and sequence content between species (11,12). The length of this region also varies between isolates of *T. brucei* and contains rapidly evolving repetitive DNA sequences (13–17). Transcription of VR sequences occurs at relatively low levels in *T. brucei* (15,17,18) and has not been detected in *L.tarentolae* (12).

The function of the VR remains obscure. It has been proposed that it may contain the origin of replication for the maxicircle (13). Maxicircles appear to replicate as network bound rolling circles (19,20) and the initiation site for the replication of the leading strand has been mapped to the VR in *C.fasciculata* (20). Cloned VR fragments from *Crithidia oncopelti* (12) and *L.tarentolae* (21) have been shown to exhibit autonomous replicating sequence (ars) activity in yeast, although the significance of these finding is unclear since sequences from elsewhere in the maxicircle also have ars activity (22). The sequence of only ~4 kb of the 12 kb *L.tarentolae* VR (23,24) and ~2.5 kb of the 7.5 kb *T. brucei* VR (15–17,24) has been previously reported. Analysis of this sequence reveals the absence of any substantial open reading frames (ORFs) but a similar structure of repetitive elements in the two species (albeit with different sequences), lending further support to a non-protein coding function.

* To whom correspondence should be addressed

+ Present address: Department of Biology, King's College, Wilkes-Barre, PA 18711, USA

We report here the complete nucleotide sequence of the VR from *T. brucei* and show that it is comprised of three sections with different sequence characteristics. Two sections contain short tandem repeats, while the third contains non-repetitive DNA, in addition to a tandem duplication. Heterogeneity between different isolates is confined to the first two (repetitive) sections, one of which contains sequences related to the minicircle replication origin and to topoisomerase II binding elements. These data suggest that the *T. brucei* VR contains the maxicircle replication origin and that replication of maxicircles and minicircles may share common features.

MATERIALS AND METHODS

Parasites and nucleic acids

The *Trypanosoma brucei brucei* stock EATRO 164 (IHRI 1)(25) was used for most studies. The dyskinetoplastic (Dk) strain derived from this stock has been described before (26). Other trypanosome isolates have been described previously (27), except for *T. b. brucei* STIB 247 (IHRI 5), *T. b. brucei* STIB 378 (IHRI 6), *T. b. gambiense* ILRAD 1375 (IHRI 13), *T. b. rhodesiense* KETRI 2473 (IHRI 20), and *T. b. rhodesiense* EATRO 1895 (IHRI 21). Bloodstream and procyclic form parasites were grown and harvested as previously described (25). Kinetoplast DNA (kDNA) and total cellular RNA were isolated using published methods (6,28,29).

Cloning and sequencing

Several strategies were used to obtain plasmid clones containing variable region sequences. Two clones (pTKR10 and pTKR9) were obtained from *Eco*RI-digested kDNA as previously described (18). Another series of clones (pA1, 3, 5, 8, 10, 15, 19, 20, 26 and 33) were obtained by *Dra*I digestion of *Bam*HI-linearized maxicircle DNA and ligation into *Dra*I-digested pBR322. *Dde*I, *Mbo*II and *Taq*I digests of the same material were treated with the Klenow fragment of DNA polymerase to produce blunt ends (30) and cloned into a *Sma*I- or *Acc*I-digested pUC13 derivative to obtain clones pD4 and p15; pM6, pM7, pM7 and pM12; and pT11, respectively. A series of clones (mA1−mA10) was obtained by partial *Dra*I digestion of the purified 4.7 kb R_4-R_1 fragment containing the 3′ portion of the VR (18) and ligation into *Sma*I- or *Sma*I+*Eco*RI-digested M13mp8 or mp9. This fragment was also digested with *Dde*I , filled with Klenow, and ligated into *Sma*I-digested pBS$^+$ to obtain clones pC7, C16, C17, C19, C20, D1, D5, K15, K16, L17, L18, M2, and M18. The remaining portion of the maxicircle was cloned by PCR amplification of kDNA using primers TbVR-1 (CCGAATTC AATCACCATTCCC-TAAG) and TbVR-2 (CCGGATCC TAACTAGTTACTTACTTCC). The location of these primers, which contained *Eco*RI or *Bam*HI sites(underlined) is indicated in Fig. 1. Thirty cycles of amplification were carried out using 1 min. denaturation at 94°C, 1 min. annealing at 45°C and 2.5 min. extension at 72°C. Following a final extension for 10 min. at 72°C, the 1.8 kb PCR product was PEG-precipitated (31), digested with *Eco*RI+*Bam*HI and ligated with *Eco*RI+*Bam*HI-digested pBluescriptII SK$^-$ and transformed into SURE™ cells.

Many of the clones described above were subcloned into plasmid and M13 vectors using conventional methods and by ExoIII digestion (Erase-a-base, Promega). Single-stranded DNA from M13 clones and double stranded DNA from plasmid clones was sequenced using the dideoxy chain termination method (32−34) or with a Taq Dye Primer Cycle Sequencing Kit on a model 373A DNA sequencer (Applied Biosystems Inc.). Sequence analysis was carried out using DNASTAR (DNASTAR Inc., Madison, WI), ESEE (35), and CLUSTALV (36) software.

Southern and Northern blot analysis

Isolated kDNA was digested with the appropriate restriction enzymes, separated by agarose gel electrophoresis, transferred to Nylon or nitrocellulose membranes, and hybridized with [^{32}P]-labelled nick translated probes, as described previously (37). Total RNA was separated on formaldehyde-agarose gels, transferred to nitrocellulose membranes and hybridized with [^{32}P]-labelled probes using previously described techniques (7).

RESULTS

We determined the complete nucleotide sequence of the variable region (VR) sequence from the *T. b. brucei* EATRO 164 (IHRI 1) maxicircle from 46 independent (43 conventional and 3 PCR) clones. The 7395 nt sequence bounded by the ND5 stop codon and the putative 5′ end of 12S rRNA is presented in Fig. 1. Partial sequences of this region have been published previously, both by ourselves (nts 1−498, 1066−1895 and 6033−7395) (15,24) and others (nts 1899−2760 and 6033−7395)(17). Differences between sequences from IHRI 1and another isolate MITat 1.1 (clone 427−60) (17) are indicated in Fig. 1. The A_2-A_3 sequence from the latter isolate contains a deletion of 19 bp (nts 2480−2496), addition of single A residues at nts 2510 and 2517, deletion of an A at nt 2428 and a T at nt 2438, and a C → G base change at nt 2737. Correction of our previously published sequence (24) now reveals only two differences (deletions at nts 6043 and 7318) between these isolates 3′ to the *Eco*RI site at nt 6033. Analysis of the sequence reveals frequent stop codons in all reading frames; the longest ORF having the potential to code for 115 amino acids. It is thus unlikely that the VR has a protein coding function.

Dot matrix analysis of the VR reveals three distinct sections (I, II and III), each of which is characterized by the presence of different repetitive sequences (see Figs. 1 and 2A). As evident from the dot matrix analysis, section I has a complex structure, containing 15 repeats of a sequence composed of two different types of sequence, which we have designated I-A and I-B (Fig. 2B, Table I). The I-A sequence (Fig. 2C) varies in size (31−187 bp) between repeats with an average length of 176 bp (excluding repeats 1, 2, and 15, which are unusual). Repeats 1and 2 are the least conserved, diverging substantially (<60% homology) in the 5′ portion of the sequence and repeat 15 contains a truncated copy of the A sequence. The homology between the remaining repeats varies between 80% (9 vs. 11 and 9 vs. 13) and 99% (12 vs. 14), and the 3′ end is more highly conserved between repeats than the 5′ end (Fig. 2C). The I-B sequence consists of variable numbers of short repeated sequences that are A-rich . These sub-repeats fall into four classes (B1-B4) whose consensus sequences are $[A_{3-6}(G)TA_{4-5}C]_{1-21}$, $[A_{4-9}(T)C]_{1-6}$, $A_{5-9}TA_{1-2}GT(/C)C(/C)A_{3-7}TT(C)]_{1-17}$, and $[A_{5-11}TA_4C$-$A_{4-6}TT]_{1-2}$, respectively (Fig. 2B). Bases shown in parenthesis occur in a significant minority of the repeats. The similarity among these four sequences indicates that they may be derived from a common ancestral sequence. As indicated in Fig. 2B and Table I, the number and order of I-A and I-B sequences varies among repeats, but in most they occur in the order A-B2-B3-B4. In addition to their differences in I-A sequence, repeats 1 and 15 contain no I-B_3 and I-B_4 sequences, while repeat 15 contains

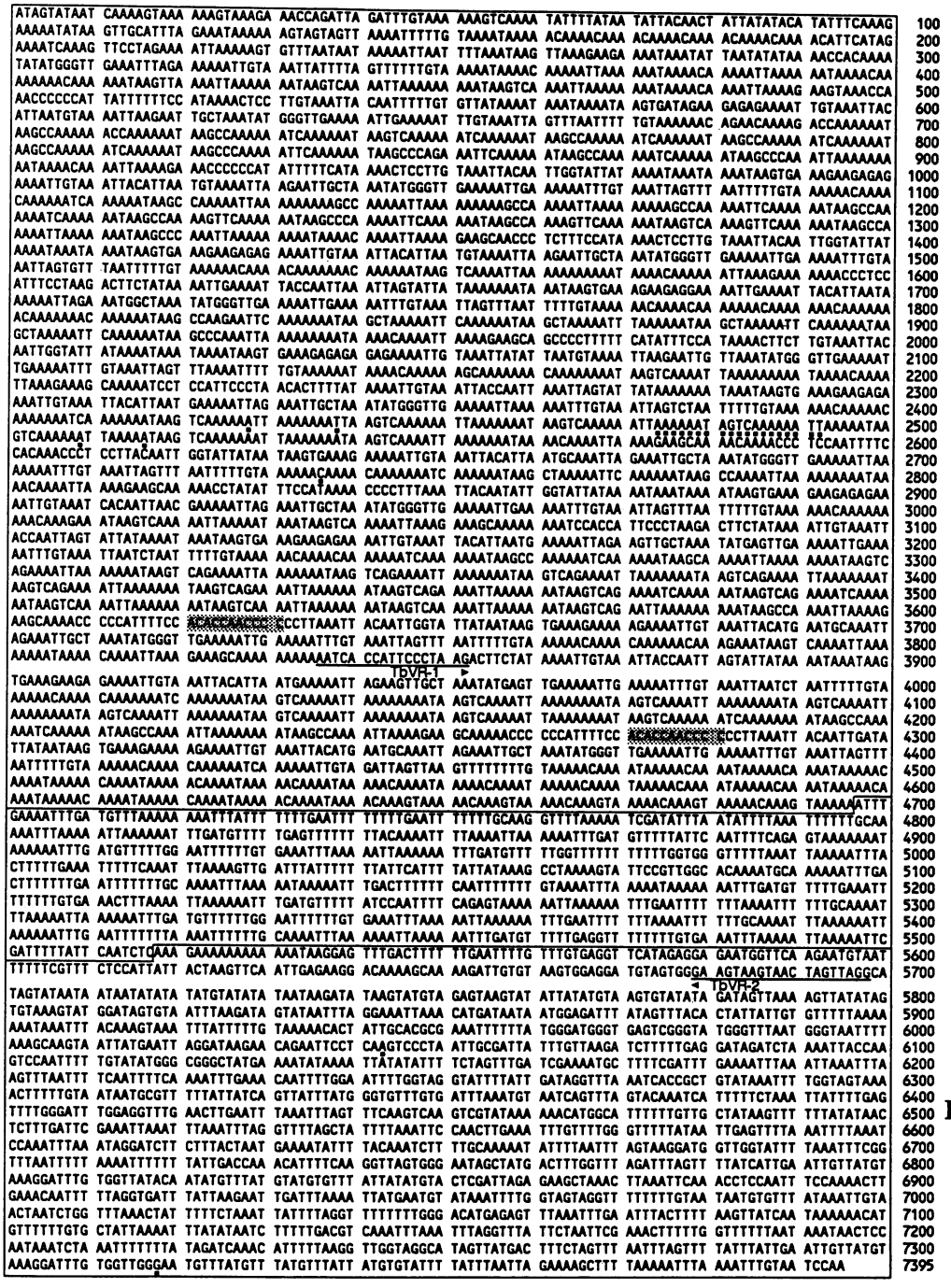
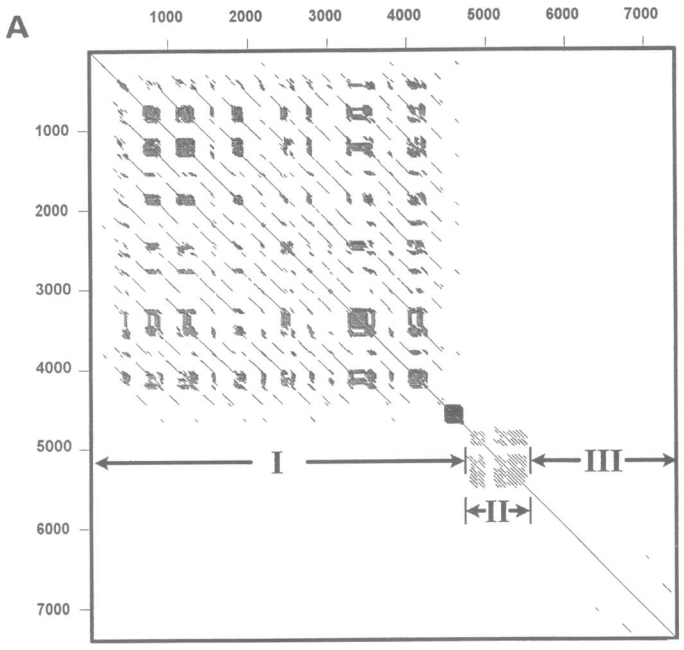
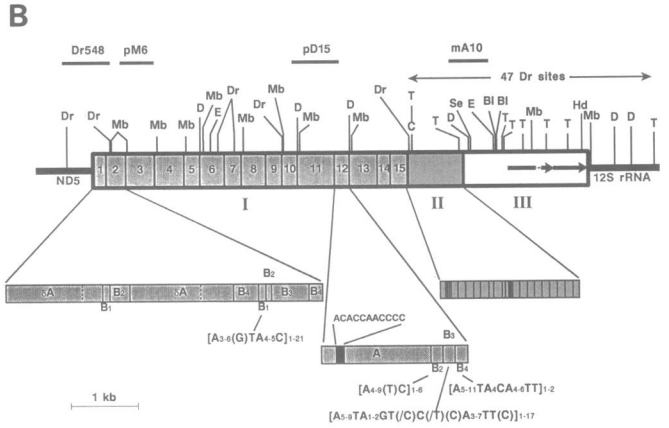


Figure 1. The nucleotide sequence of the variable region (VR) from *T. b. brucei* EATRO 164 (IHRI 1). The 5′ end of the sequence begins immediately 3′ to the TAA stop codon of ND5 and its 3′ end is adjacent to the 5′ end of the 12S rRNA gene. The three sections (I, II and III) identified in the text are boxed. Nucleotides which differ from the published sequence for MITat 1.1 (17) are indicated by dots underneath. The oligonucleotide primers (TbVR1 and TbVR2) used for PCR amplification of a portion of the VR are indicated by underlining. The two copies of the 11-mer sequence with homology to the minicircle conserved sequence (38,39) are shaded. This sequence was submitted to the EMBL database and assigned accession number Z15118.

21 copies of the I-B_1 sequence, which occurs only infrequently in other repeat units (see Table I).

The sequence immediately 3′ to repeat I-15 (section II, Figs. 1 and 2B) has a different character from section I. It is composed largely of oligo(dT) and oligo(dA) tracts that occur as tandem repeats of the sequence [$A_3T_3A_5TTA_6T_3GATGT_5$ $G/AA_2T_7GYA(A)$]. Sequence conservation between the 17 complete copies of the repeat ranges from 70−89%, with greater conservation at the 5′ end than at the 3′ end (Fig. 2D). In addition, there are 7 incomplete copies of the sequences, which fall into two classes, one (Δ2, Δ3, Δ4, Δ14, and Δ15) of which contains only the 3′ portion of the consensus sequence and another (δ12, and δ13) which contains the 5′ portion of the consensus. The remainder of the VR sequence (Section III, Figs. 1 and 2B)





C

Repeat	Coordinates
1	0001-0151
2	0193-0348
3	0486-0673
4	0915-1089
5	1347-1520
6	1584-1766
7	1951-2133
8	2203-2387
9	2570-2729
10	2810-2987
11	3047-3226
12	3596-3759
13	3818-3999
14	4234-4409
15	4430-4461
Consensus	

VR I-A Repeat Sequences

```
                                                    .T.GTATAATCAAA.GTAA.AA.G.AAAGAAACC.GAT..G.T.TGTA...AAGT
                                                    ATTC...G..AAT.AAAGT.CCTAG.AA.T.AAAAAGT..T.A....T.A....T..T.TT
....... .T...CCAAA   ...C    .....A..TTTT..          .......T..  ..........  ....   TTT..G.................A...
.......A             ...C    .... A..TTT  .          .......T..  ..........  ....   ......................A...
....... ....         ... T   . .. .    ..            .......T..  ..........  ....   ......................A...
... ...AAA..         ... T   ..... .   .. TA.G..TTCT.....       ...A.......C....AAATTA..........  .......  ...
....... ...G         ...CTTTTT... A..T   ..          .......TT.  ..........  ....   ......................A...
... ...A......A      T.. T   .....       ..CTA.C..TTTT.....      ..........C....AAATTA..........  .......  ...
....... ......CAAAA  ... T   ...  A..TT  ..   .C.A        ..C.TCC  T.    .. ....   ..........              ...
....... ......A      .. T    .. A..T    ..                 .......C. C .. ........  ...A   TT.....................A...
... ...A......AAAA   T..     A.....      ..CTA.G..TTCT.....      ..........C....A   ..............              ...
....... ......       ...     .....  ..   ..  .C..  C    ...C..CC..AA.TTAC.A  ..    ..........              ...
... ...A......AAAAAAAT.      A.....      ..CTA.G..TTCT.....      ..........C....A   ..............              ...
....... ......A      ...C    .....  ..   ..  .C..  C    ...C..CC..AA.TTAC.A  ..    .A........              ...
Consensus
AAAAGAA GCAAAA       CCC     CCATT TT   CC  ASAC    ATAAAACYCC  TTGTAAATTAC AATT     GGTATTATAAAATAAATAAA TAA
```

```
1   CAA..TATTTT.T.AT.TTACAACT....   .AT.CAT.TTTCA.AG..AAA ..  ..A  ....C . T..A....T.AAAAA...  G.....   ..AA  .......
2   AAAT...TT.A...AG...A.AA.T....A  .AT..AA..  C CAC..AA  ..  ..... ..... T..A......A  ..........  ...T . G  T.......
3   ....T...  ..........  ......T.......AG..  ..........  ......  ..............  ..  .......
4   ........  ..........  ......T.......AG..  ..........  ......A  ..............  ..  .......
5   ........  ..........  ......T.......AG..  ..........  ......  ..............G.TT..  .......
6   ........  ....  ......A.........  ....AA..  ..........G..  ......  ..............  ..  .......
7   ........G..........T.  ......T.......AG.....T..  ..........  ......  ..............  ..AA  .......
8   ........  ..........  ......A.........  ..........  ..........A.....  ..........C..  ..  .......
9   ........  ..........  ......  ..........  ..........  ..........A....A  ..........  ..  .......
10  ........  ..........C...A....C.A.  ..........  ..........  ......  ..............  ..  .......
11  ........  ..........  ......A.........G..........A.........  ......  ..............A.C.  ..  .......
12  ........ . .A..........  .G....C...  ..........  ..........  ......  ..............  ..  .......
13  ........  ..........  ......A.........G..........A.........  ......  ..............A.C.  ..  .......
14  ........ . .A..........  .G....C...  ..........  ..........  ......  ..............  ..  .......
15                                                                A.....  .....G.......  ..GTTT.....
Consensus
GTGAAAGA AGAGAGAAAATTGTAAATTACA TTAATGNAAAATTAGAAATTGCTAAATATGGGTTGAAAAATT GAAAAA  TTTGTAAATTAGTTT  AA   TTTTTGT
```

D

Repeat	Coordinates
1	4697-4727
Δ2	4728-4737
Δ3	4738-4748
Δ4	4749-4759
5	4760-4800
6	4801-4845
7	4846-4892
8	4893-4931
9	4932-4981
10	4982-5016
11	5017-5063
δ12	5064-5080
δ13	5081-5101
Δ14	5102-5111
Δ15	5112-5121
16	5122-5163
17	5164-5209
18	5210-5251
19	5252-5296
20	5297-5341
21	5342-5386
22	5387-5432
23	5433-5479
24	5480-5517
consensus	

VR II Repeat Sequences

```
1   ....G          ..............AAAA..AA....A
Δ2                                 .......   . .A
Δ3                                 .......T  . .A
Δ4                                 .......   .C.A
5   GGT........ .CG.T.  ...A..A....  A.. ......   .C.A
6   .......... ..... G.G .......TA   C.
7   .......... ..... A  ..CAA..TTCA .AGT
8   ...  .... . ..... GGAA.....      . TG
9   .......... ..... GGTT.....       .GTGGG
10    ..TT... ..... ... .C  ....      .A.A TTTTTC
11  .........G ..G.T  ..  ..T....ATTCA....A..ATAAA.CCT
    ....G . .TTCCG..G.C
.C.  ...TG  C...........C
                                   .......   . .A
                                   .......   .C.
.....  .....CT.....CA.  .......   .T.
.....  .A..................GA..  .......T   .TG
...C..T....  ..................ATC CAA....   C
.G.G.  ..................AT.....TA..A.....   .C.
.....  ..................GGAA.....          .TG
.....  .....AT.....TA..A.....               .C.
.....  .....AT.....TA..A.....               .C.
.....A..............AT.....TA..A.....       .C.
.....A..................GAGG .......T       .TG
.....  ..... ..C... ....  A  ..CAA.CTC
AAATTTAAAAA TTAAAAAATTTGATGTTTTT RAA TTTTTTT    GYA
```

E Repeat — **VR III Repeat Sequences**

```
1  TTTTCAAAATTTGAAACAATTTTGGAATTTTGGTAGGTATTTTATTGATAGGTTT    AAATCACCGCTGTATAAATTTTGGTAGTAAAACTTTTT-GTAATAATGCGTTTTTATTATCAGTTATTTATGGGTGTTTGT
2  ...C.....C.............          .....GA......A.G.AT.GATTTA....T.TGAA.....................GGTT.....T.........T....A. .AAT

1  GATTTAAATGTAATCAGTTTAGTACAAATCATTTTTCTAAATTATTTTGAGTTTTGGGATTTGGAGGTTTGAACTTGAATTTAAATTTAGTTTCAAGTCAAGTCGTATAAAAAACATGGCATTTTTTGTTGCTATAAGTT
2          .....CT.A.C.G..TT...CT.....................AG.....TTTTG  ...CA.GA..GT..A.....G......C...T....T.TCA  .............  ........ .....T.AAA

1  TTTTATATAACTCTTTGATTCGAAATTAAATTTAAATTTAGGTTTTAGCTATTTTAAATTCCAACTTGAAATTTGTTTTGGGTTTTTATAATTGAGTTTTAAATTTTAAATCCAAATTTAAATAGGATCTTCTTTACTAA
2  ...ATATA.T..T.....CGTC.....      ..............      ...C...T.CGA.....          .....T......A          .....AACTCCAAT....C.....

1  TGAAAATATTTTACAAATCTTTTGCAAAAATATTTTAATTTAGTAAGGATGGTTGGTATTTTAAATTTCGGTTTAATTTTTAAAATTTTTTTTATTGACCAAACATTTTCAAGGTTAGTGGGAATAGCTATGACTTTG  G
2                                                                                                                                  .....A..T..........T......G..A..C....T.........CTA.

1  TTTAGATTTAGTTTTATCATTGAATTGTTATGTAAAGGATTTGTGGTTATACAATATGTTTATGTATGTGTTTATTATATGTACTCGATT     AGAGAAGCT     AAACTTAAATTCAAACCTCCAA
2  ....AT..AGT...AT.T.............................GGGA  ..........T.A..........G....T.TT...TAATT...A.....TTTAA...T.....A.TTGTAA.....
```

F

```
|←— I-2B3-1 —→|←— I-2B3-2 —→|←— I-2B3-3 —→|← I-2B2-1
407- AAAAAATAAGTTAAAATTAAAAAAATAAGTCAAAATTAAAAAAAAATAAGTCAAAATTAAAAAAA - 474
      ||::| |||:| :::   ||::| |||:| :::   ||::| |||:| :::
      GTNWAYATTNATNNR   GTNWAYATTNATNNR   GTNWAYATTNATNNR
      topoII box        topoII box        topoII box
```

is characterized by a more even base composition and absence of short tandem repeats. However, the 3′ portion of this region does contain two tandem copies of a longer (545−700 nt) sequence (Figs 1 and 2B). These repeats occur immediately upstream of the 12S rRNA gene. Alignment of the two copies (Fig. 2E) shows considerable divergence (73% identity), with several additions and deletions. Most noticeably, the more 3′ copy lacks several regions present in the 5′ copy.

Some I-A repeats contain a short sequence with homology to the putative minicircle replication origin. The sequence of the 5′ portion of repeat I-A is not highly conserved between copies, but all (except 1, 2, and 15) contain oligo(dC) tracts (Fig. 2C). Repeats 12 and 14 contain the sequence ACACCAACCCC (see Fig. 1), which is identical to 11 nt of the 12 nt sequence (TACACCAACCCC) that is conserved in all trypanosomatid minicircles (38,39). This region is the site of a small gap in replicating minicircles (38) and has been proposed to have a role in minicircle replication (19,39). It is possible that these sequences serve a similar function in the maxicircle.

Replication of maxicircles (and minicircles) has been shown to involve topoisomerase II (19). The consensus sequence of repeat unit I-B_3 ($A_{5-9}TA_{1-2}GT(/C)C(/C)A_{3-7}TT(C)$, Fig. 2B) shows striking similarity to the consensus topoisomerase II cleavage site (GTNA/TAT/CATTNATNNG/A) from *Drosophila* (40). Alignment of this topo II box with I-B_3 repeats from repeat I-2 shows matches at 13/15 positions and 7/8 of the conserved nucleotides (Fig. 2F). Two other sequences associated with the topo II box in Scaffold Associated Regions (SARs) in *Drosophila* are the A-box (AATAAAT/CAAA) and T-box (TTA/TTT/ATTT/ATT)(40). These sequences are reminiscent of the I-B_1 (AAAATAAAAAC or GTTTTTATTTT, in the reverse complement) and I-B_2 (AAAAAAC) repeats, suggesting a possible role for the B repeats in binding topoisomerase II.

The size of the VR differs considerably among isolates of *T. brucei*. In order to localize the sequence(s) responsible for this variation, we carried out Southern blot analyses of kDNA from several different isolates using probes derived from different regions of the VR. A probe (Dr548) derived from the 5′ boundary of the VR (see Fig. 2B) hybridized with a ~0.6 kb *Dra*I fragment in 12 different isolates (Fig. 3A and data not shown). This fragment is of the size predicted (548 bp) from the IHRI 1 sequence, indicating that all isolates contained similar sequence 5′ to the *Dra*I site within repeat I-2 (see Fig. 2B). Hybridization of restriction enzyme-digested kDNA with probe mA10 (see Fig. 2B) helped to define the 3′ boundary of sequence divergence among isolates. This probe detected a 2.2 kb *Dde*I fragment in all isolates (Fig. 3B and data not shown). The size of this fragment is consistent with that (2681 bp) predicted from the IHRI 1 sequence, and is reduced in all isolates to 0.4 kb, as predicted from the sequence, by double digestion with *Dde*I+*Eco*RI. Thus, the region 3′ to the *Dde*I site at the beginning of section III (see Fig. 2B) is conserved in all isolates. The mA10 probe also detects a second band of 1.7 kb in IHRI 1, IHRI 5, IHRI 21 (Fig. 3B), IHRI 3 and IHRI 4 (data not shown) which corresponds to the *Dde*I fragment of 1774 bp (predicted from the IHRI 1 sequence) extending 5′ through section II to repeat I-12. However, this band was replaced in other isolates by a less intense band which varied in size between 1.5 kb (IHRI 9) and 1.9 kb (IHRI 13 and IHRI 17)(Fig. 3B and data not shown). These data and similar results obtained with *Taq*I digests (data not shown), indicate that the intra-specific size variations occur 5′ to the *Dde*I site (at nt 5623) in section III.

A probe (pM6) covering repeat I-3 (see Fig. 2B) hybridized with different sized *Dra*I fragments in different isolates (Fig. 3C), confirming that intra-specific size variation occurred within section I. A 1.6 kb *Dra*I fragment which was reduced slightly by subsequent *Alu*I digestion was seen in IHRI 1, IHRI 5, IHRI 21 (Fig. 3C), IHRI 3 and IHRI 4 (data not shown). This corresponds to the 1618 bp *Dra*I fragment spanning repeats I-2 to I-6 predicted from the IHRI 1 sequence (see Fig. 2B). A larger *Dra*I fragment hybridized in IHRI 6 (~4.7 kb), IHRI 20 (~4.1 kb), IHRI 22 (~4.1 kb), IHRI 13 (~2.7 kb)(Fig. 3B), IHRI 9 (~3.8 kb), IHRI 16 (2.2 kb) and IHRI 17 (2.5 kb) (data not shown). Double digestion with *Dra*I+*Alu*I digestion resulted in one to four smaller (0.36−1.3 kb) fragments in these isolates. Similar results obtained with other enzyme digests and probes from this region (data not shown) indicated that the strain-specific variation in restriction fragment length was not due solely to abolition or addition of sites by small base substitutions, but was caused by addition or deletion of sequence between the *Dra*I site in repeat I-2 (at nt 253) and the *Dde*I site in the 5′ end of section III (at nt 5623). The results of these and other (unpublished) experiments are summarized in Fig. 4, which shows the variation in the length of the VR in different isolates. As indicated in this figure, the differences in the size of maxicircles between isolates is accounted for by variation in the length of sections I and II of the VR. Comparison of IHRI 1 VR sequence with partial sequence from MITat 1.1 (17) indicates that the latter has one fewer copy of the I-B_3 sequence in repeat I-8 (nts 2480−2496, Fig. 1). These sequences are otherwise almost identical, and are located in the same position of the VR, indicating that the major sequence difference between isolates is likely a difference in the number of copies of the I-A, I-B and II sub-repeat units.

Northern blot analyses were used to identify transcripts from the VR. A probe (pD15) from section I (see Fig. 2B) hybridized to a heterogeneous smear of transcripts, ranging in size from 0.2

Figure 2. Sequence organization of the variable region. **A.** The VR sequence compared to itself using the DNASTAR Dotplot program with an 80% minimum match over of window of 30 nucleotides, and a minimum quality of 10. The locations of sections I, II and III are indicated. **B.** Diagrammatic representation of the VR organization in *T. b. brucei* EATRO 164. The size and location of each of the repeats (1−15) within section I are shown; and the sub-repeat structures of repeats I-1, I-2, and I-12 are indicated by the expanded lower bars. The dotted lines within the δA sub-repeats of repeats I-1 and I-2 (lower left) indicate the approximate 5′ boundary of homology with the other A repeats. The consensus sequences of the I-B sub-repeats are shown and the solid black bar shows the location of the conserved 11 nt sequence (ACACCAACCCC) within repeat I-12 (lower middle). The lower right expanded bar diagrams the size and location of the section II repeats (1−24). The arrows in section III indicate tandemly duplicated sequences. The locations of probes used in Southern and Northern blot analyses are indicated. The following abbreviations for restriction enzymes are used: Bl, *Bgl*II; C, *Cla*I; D, *Dde*I; Dr, *Dra*I; E, *Eco*RI; Mb, *Mbo*II; Se, *Spe*I; and T, *Taq*I. **C.** The consensus sequence of the A repeats from section I and alignment of individual sequences. The coordinates of each repeat within the VR are shown and identities to the consensus sequence are indicated by dots. **D.** Alignment of section II repeats with their consensus sequence. **E.** Alignment of the duplicated sequence from section III. The upper and lower sequences have coordinates 6215−6888 and 6889−7395, respectively, within the VR. **F.** Alignment of the three B_3 sub-repeats from repeat I-2 with the topoisomerase II cleavage site consensus sequence (40). The | indicates identity to a strictly conserved nucleotide and : indicates a match with an ambiguous nucleotide. The coordinates within the VR are indicated at either end of the sequence.

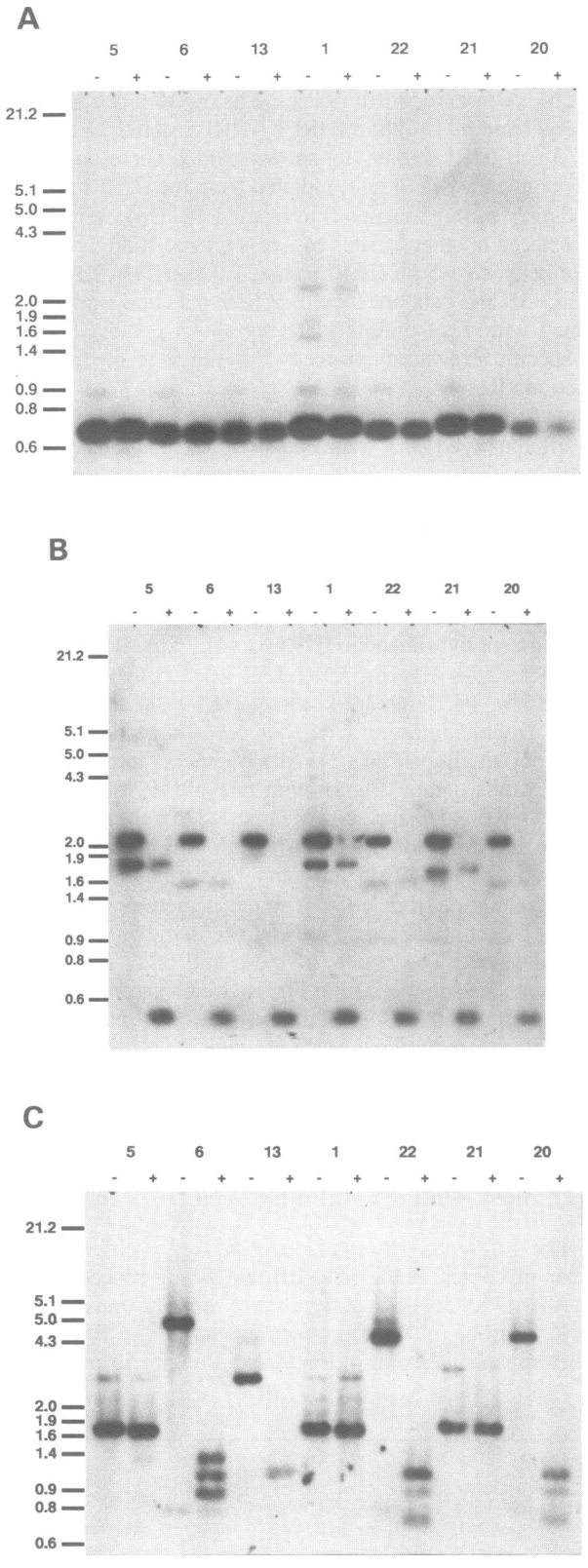


Figure 3. Localization of intra-specific size differences in the variable region. kDNA (1 μg) from IHRI 5, 6, 13, 1, 22, 21 and 20 was digested with *Dra*I with (+) or without (−) *Alu*I (panels A and C) or *Dde*I with (−) or without (+) *Eco*RI (panel B), fractionated by agarose gel electrophoresis, and probed with Dr548 (panel A), mA10 (panel B) or pM6 (panel C). Markers derived from *Eco*RI+*Hin*dIII digestion of λ DNA are shown in kb.

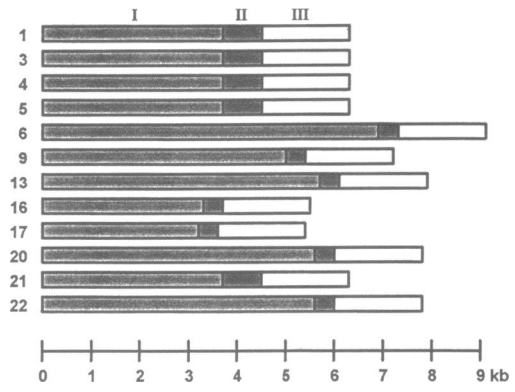


Figure 4. Diagram of intra-specific size differences within the variable region. Changes in the size of the different sections of the variable regions are shown schematically for 12 different isolates of *T. b. brucei* (IHRI 1, 3, 4, 5, 6, and 9), *T. b. gambiense* (IHRI 13, 16 and 17) or *T. b. rhodesiense* (IHRI 20, 21, and 22).

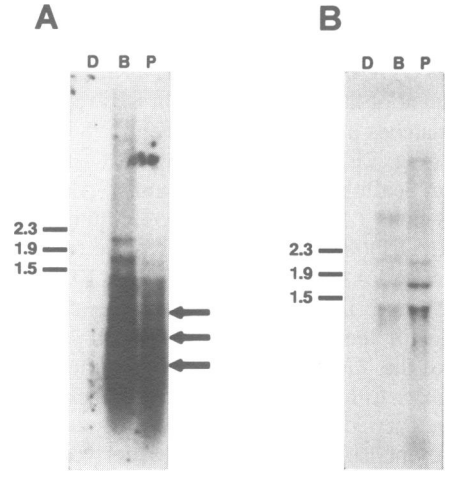


Figure 5. Northern blot analysis of VR transcripts. Total RNA (10 μg) from dyskinetoplastic (D), bloodstream (B), and procyclic (P) forms of *T. b. brucei* EATRO 164 was separated by denaturing agarose gel electrophoresis and hybridized with probed pD15 (panel A) or mA10 (panel B). Transcripts of 960, 780, and 570 nt detected with pD15 are indicated by arrows. The locations of the three cytoplasmic rRNA bands are indicated by their sizes (in kb).

kb to >5.0 kb, in total RNA from both bloodstream (Fig. 5A, lane 2) and procyclic (lane 3) forms. In agreement with previous results (15,18), discrete transcripts of ~570, 780 and 960 nt were detected in bloodstream and procyclic forms, although their abundance differed between life-cycle stages. Three apparently larger bands were also detected, but probably represent artifacts due to shadowing from rRNA. No hybridization was seen with RNA from dyskinetoplastic trypanosomes (which lack kDNA)(26), confirming that the signal is not due to hybridization with nuclear transcripts. DNase treatment of the RNA did not affect the hybridization signal (data not shown), eliminating the possibility that it was due to DNA contamination. Similar results were obtained with other probes from this region (data not shown). Hybridization with a more 3′ (and single-stranded) probe (mA10) also showed a smear (and rRNA shadowing) (Fig. 5B). However, there appeared to be differences in the hybridization

Table I. Organization of I-A and I-B repeats from the *T. brucei* EATRO 164 Variable Region

repeat	I-A	I-B_1	I-B_2	I-B_3	I-B_4	coordinates
1	1*	1	6	0	0	0001–0192
2	1*	1	1	3	2†+1	0193–0485
3	1	0	3	10	1	0486–0914
4	1	0	3	11	1	0915–1346
5	1	0	3	1	1	1347–1583
6	1	0	6	6	1	1584–1950
7	1	1	2	1	1	1951–2202
8	1	0	3	7	1	2203–2569
9	1	0	3	2	1	2570–2809
10	1	0	2	2	0	2810–3046
11	1	0	3	17	0	3047–3595
12	1	0	2	1	1	3596–3817
13	1	0	3	11	0	3818–4243
14	1	0	3	0	0	4244–4429
15	1**	21	1‡	0	0	4430–4696
total	15	24	44	72	11	

*Limited homology to consensus.
**Truncated at 5′ end.
†Two copies of I-B_4 occur between I-A and I-B_2.
‡I-B_2 occurs between I-A and I-B_1.

pattern between bloodstream and procyclic forms, and hybridization with poly(A^+) RNA suggested possible differences in polyadenylation between life-cycle stages (data not shown). Hybridization with a probe from the other strand showed similar results (data not shown) indicating that both strands are transcribed in this region. Further clarification of the nature of the transcripts from the VR requires more experimentation.

DISCUSSION

The region flanked by the ND5 and 12S rRNA genes from the *T. brucei* maxicircle varies in size between different isolates (13–17) and diverges in sequence among trypanosomatid species (11,12). We have determined the nucleotide sequence of this entire variable region from *T. b. brucei* EATRO 164 (IHRI 1), thus completing the entire maxicircle sequence from this organism. The 7395 bp VR sequence is largely composed of different repetitive sequences which divide the region into three distinct sections. The most 5′ section (I) is composed of 15 tandem repeats of a unit comprised of two sub-sections (I-A and I-B). The I-A sequence occurs once in each repeat unit, is ~176 bp long and is generally well conserved (80–99%) between copies, although only partial or divergent copies are present in three repeat units at the ends of section I. The remainder of each section I repeat is composed of a variable number of related A-rich repeats (I-B_1, I-B_2, I-B_3, and I-B_4). Section II consists of 17 complete and 7 partial copies of a repeat containing alternating oligo(dA) and oligo(dT) tracts. Section III contains 693 bp of non-repetitive sequence, upstream of a long tandemly duplicated sequence. The two copies are somewhat divergent (73% identity); the more 3′ sequence lacking ~150 bp relative to the other. The definition of these repeat units differs somewhat from those previously identified (16). The element previously referred to as A corresponds to I-A+I-B_2, B corresponds to I-B_3, C corresponds to I-B_4+I-A, and D corresponds to repeat II. In addition, the tandem duplication in section III was not described in previous publications of sequence from this region, known variously as R_1T_4 (41), TBDV3 (24) and R_3D_3 (17). Southern blot analysis of kDNA from several different isolates of *T. b. brucei*, *T. b. gambiense* and *T. b. rhodesiense* indicates that the variation in the size of the VR between isolates is limited to sections I and II. Comparison with the limited sequence available from other isolates suggests that the variation may be due to differences in the number of repeat and sub-repeat units within these sections, as suggested by earlier by others (13). The size differences observed in the VR among different isolates may prove useful for epidemiological studies. Southern blot analysis of the appropriate restriction digests of parasite DNA using probes from sections I and II of the VR would discriminate between different strains of *T. b. brucei*, *T. b. gambiense* and *T. b. rhodesiense*.

The availability of the complete sequence of the *T. brucei* maxicircle sheds some new light on possible functions for VR. The absence of any significant ORFs suggests a role other than coding for proteins. Other eukaryotic mitochondrial genomes contain non-coding regions of various sizes (125 bp–20 kb) which often contain tandemly arranged, repeated sequences. These regions have, in some cases, been shown to contain signals for the initiation of replication and transcription (42). It has been proposed that the VR of the trypanosomatid maxicircle has a similar function (13). Maxicircle replication has been reported to occur by a rolling circle mechanism while still attached to the network (see 19 for review). DNA synthesis of the leading strand appears to initiate within the VR (20). The ends of the resultant linear maxicircle molecule are thought to contain repeated sequences, another hallmark of the VR. The presence of a replication origin within the VR is strongly supported by the finding that two of the I-A repeats contain the sequence ACA-CCAACCCC, which is also found at the putative replication origin of trypanosomatid minicircles (38,39). The 11 nt maxicircle sequence forms part of the 13 nt sequence (TTACA-CCAACCCC) conserved in all *T. brucei* minicircles, which is the site of a small gap in replicating minicircles (38). This region appears to be the site of leading strand DNA synthesis primed by a small RNA transcript (19). It is perhaps surprising that the conserved minicircle sequence should be found in the maxicircle, since the two molecules are thought to replicate by different mechanisms. Minicircle replication is thought to involve a Cairns mechanism after detachment from the kDNA network, in contrast to the rolling circle mechanism for the maxicircle (19) . However, it is possible that the shared sequence reflects some commonalty between these mechanisms. One possibility is that this sequence serves as a common signal for processing of RNA transcripts to serve as replication primers, perhaps in a manner analogous to that for the site-specific ribonucleoprotein endoribonuclease (RNase MRP) purported to be involved in replication priming of mammalian and yeast mitochondrial DNA (43–45). Thus, a shared activity may suffice to begin replication of both maxicircle and minicircle DNAs, with subsequent steps differing between the two. The VR sequence does not appear to contain the ACGCCC sequence involved in initiation of DNA synthesis of the lagging strand of minicircles (19), suggesting a different mechanism is used for this step.

In addition to a potential role in initiation of DNA synthesis the repeat sequences of section I may also be involved in binding topoisomerase II, an enzyme known to be involved in replication of the kDNA network. Repeats I-B_3, I-B_1 and I-B_2 show striking similarity to the consensus topoisomerase II cleavage site, A-box, and T-box sequences associated with binding of *Drosophila* DNA to the nuclear scaffold (40). Perhaps these sequences serve similar

roles in *T. brucei* and may be involved in maintenance of kDNA network structure.

The function(s) of the repetitive sequences in sections II and III and the non-repetitive sequence in section III remains obscure. The reiterative nature of the section II repeats and their potential for secondary structure formation (due to the interspersed A- and T-tracts), suggest that this region may serve as a protein binding site. The recent finding that the 12S and 9S rRNAs are processed from a precursor transcript which extends at least 1200 nt upstream of the 12S rRNA gene (46) suggests that the promoter for synthesis of this strand may lie within the VR. The detection of heterogeneously sized transcripts from most of the VR (see Fig. 5A and B), and the apparent absence of transcripts from section II (data not shown) suggest a possible promoter function for this region. However, resolution of this question awaits further study.

The function of VR transcripts is unknown. They appear not to be tRNAs, which are nuclearly encoded (3) and sequence analysis indicates that the *T. brucei* VR does not encode gRNAs. The ND7 and CYb gRNA which are found at the 5′ and 3′ boundaries of the VR in *L.tarentolae* (10) and *C.fasciculata* (47) are not conserved in *T. brucei*, the latter being found in minicircles (unpublished data). Similarly, the absence of significant ORFs suggest they are unlikely to serve as mRNAs. It is possible that the VR transcripts represent no more than processing products from primary transcripts which are initiated within this region.

ACKNOWLEDGMENTS

We would like to thank Elke Gobright, Andrea Perrollaz and Jeanette Bertles for their excellent technical assistance. This work was supported by PHS grant NIH AI14102 to KS, who is also a Burroughs Wellcome Scholar in Molecular Parasitology.

REFERENCES

1. Stuart,K. (1983) Mol. Biochem. Parasitol., **9**, 93–104.
2. Hajduk,S., Adler,B., Bertrand,K., Fearon,K., Hager,K., Hancock,K., Harris,M., Le Blanc,A., Moore,R., Pollard,V., Priest,J. and Wood,Z. (1992) Am. J. Med. Sci., **303**, 258–270.
3. Hancock,K. and Hajduk,S.L. (1990) J. Biol. Chem., **265**, 19208–19215.
4. Read,L.K., Myler,P.J. and Stuart,K. (1992) J. Biol. Chem., **267**, 1123–1128.
5. Feagin,J.E. and Stuart,K. (1985) Proc. Natl. Acad. Sci. USA, **82**, 3380–3384.
6. Feagin,J.E., Jasmer,D.P. and Stuart,K. (1985) Nucl. Acid. Res., **13**, 4577–4596.
7. Jasmer,D.J., Feagin,J.E. and Stuart,K. (1985) Mol. Cell. Biol., **5**, 3041–3047.
8. Benne,R. (1990) Trends Genet., **6**, 177–181.
9. Stuart,K. (1991) Annu. Rev. Microbiol., **45**, 327–344.
10. Blum,B., Bakalara,N. and Simpson,L. (1990) Cell, **60**, 189–198.
11. Muhich,M.L., Simpson,L. and Simpson,A.M. (1983) Proc. Natl. Acad. Sci. USA, **80**, 4060–4064.
12. Maslov,D.A., Kolesnikov,A.A. and Zaitseva,G.N. (1984) Mol. Biochem. Parasitol., **12**, 351–364.
13. Borst,P., Weijers,P.J. and Brakenhoff,G.J. (1982) Biochim. Biophys. Acta, **699**, 272–280.
14. Borst,P., Fase-Fowler,F., Hoeijmakers,J.H. and Frasch,A.C. (1980) Biochim. Biophys. Acta, **610**, 197–210.
15. Stuart,K., Feagin,J.E. and Jasmer,D.P. (1985) In Calender,R. and Gold,L. (eds.), Sequence specificity in transcription and translation. Alan R. Liss, Inc., New York, pp. 621–631.
16. Stuart,K., Feagin,J.E. and Jasmer,D.P. (1987) In Agabian,N., Goodman,H. and Nogueira,N. (eds.), Molecular strategies of parasite invasion. Alan R. Liss, Inc., New York, pp. 145–155.
17. de Vries,B.F., Mulder,E., Brakenhoff,J.P., Sloof,P. and Benne,R. (1988) Mol. Biochem. Parasitol., **27**, 71–82.
18. Stuart,K.D. and Gelvin,S.B. (1982) Mol. Cell. Biol., **2**, 845–852.
19. Ryan,K.A., Shapiro,T.A., Rauch,C.A. and Englund,P.T. (1988) Annu. Rev. Microbiol., **42**, 339–358.
20. Hajduk,S.L., Klein,V.A. and Englund,P.T. (1984) Cell, **36**, 483–492.
21. Hughes,D., Simpson,L., Kayne,P.S. and Neckelmann,N. (1984) Mol. Biochem. Parasitol., **13**, 263–275.
22. Kim,R. and Ray,D.S. (1985) Gene, **40**, 291–299.
23. Muhich,M.L., Neckelmann,N. and Simpson,L. (1985) Nucl. Acid. Res., **13**, 3241–3259.
24. Simpson,L., Neckelmann,N., de la Cruz,V.F., Simpson,A.M., Feagin,J.E., Jasmer,D.P. and Stuart,K. (1987) J. Biol. Chem., **262**, 6182–6196.
25. Stuart,K., Gobright,E., Jenni,L., Milhausen,M., Thomashow,L. and Agabian,N. (1984) J. Parasitol., **70**, 747–754.
26. Stuart,K. and Gelvin,S.R. (1980) Am. J. Trop. Med. Hyg., **29**, 1075–1081.
27. Jasmer,D.J., Feagin,J.E., Payne,M. and Stuart,K. (1987) Mol. Biochem. Parasitol., **22**, 259–272.
28. Stuart,K. (1979) Plasmid, **2**, 520–528.
29. Johnson,B.J., Hill,G.C., Fox,T.D. and Stuart,K. (1982) Mol. Biochem. Parasitol., **5**, 381–390.
30. Sambrook,J., Fritsch,T. and Maniatis,T. (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
31. Kusukawa,N., Uemori,T., Asada,K. and Kato,I. (1990) BioTech., **9**, 66–72.
32. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc. Natl. Acad. Sci. USA, **74**, 5463–5467.
33. Chen,E.J. and Seeburg,P.H. (1985) DNA, **4**, 165–170.
34. Tabor,S. and Richardson,C.C. (1987) Proc. Natl. Acad. Sci. USA, **84**, 4767–4771.
35. Cabot,E.L. and Beckenbach,A.T. (1989) Comput. Appl. Biosci., **5**, 233–234.
36. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) Comput. Appl. Biosci., **8**, 189–191.
37. Myler,P.J., Allison,J., Agabian,N. and Stuart,K. (1984) Cell, **39**, 203–211.
38. Ntambi,J.M. and Englund,P.T. (1985) J. Biol. Chem., **260**, 5574–5579.
39. Ray,D.S. (1989) Mol. Cell. Biol., **9**, 1365–1367.
40. Gasser,S.M. and Laemmli,U.K. (1987) Trends Genet., **3**, 16–22.
41. Stuart,K. (1983) J. Cell. Biochem., **23**, 13–26.
42. Gray,M.W. (1989) Annu. Rev. Cell Biol., **5**, 25–50.
43. Chang,D.D. and Clayton,D.A. (1987) EMBO J., **6**, 409–417.
44. Stohl,L.L. and Clayton,D.A. (1992) Mol. Cell. Biol., **12**, 2561–2569.
45. Topper,J.N., Bennett,J,L. and Clayton,D.A. (1992) Cell, **70**, 16–20.
46. Michelotti,E.F., Harris,M.E., Alder,B., Torri,A.F. and Hajduk,S.L. (1992) Mol. Biochem. Parasitol., **54**, 31–42.
47. Van der Spek,H., Arts,G.-J., Zwaal,R.R., Van den Burg,J., Sloof,P. and Benne,R. (1991) EMBO J., **10**, 1217–1224.