(TAA)_n within sequences flanking several intrachromosomal variant surface glycoprotein genes in *Trypanosoma brucei*

Robert Aline Jr., Gene MacDonald, Elizabeth Brown, Janette Allison, Peter Myler, Victoria Rothwell and Kenneth Stuart

Issaquah Health Research Institute, 1595 N.W. Gilman Blvd., Issaquah, WA 98027, USA

Received 19 February 1985; Revised and Accepted 12 April 1985

ABSTRACT

In *Trypanosoma brucei* telomeric copies of intrachromosomal variant surface glycoprotein (VSG) genes are produced through a gene conversion mechanism and are expressed if the telomere is or becomes transcriptionally activated. We have analyzed a sequence that occurs 1 to 1.5 kb 5' to several intrachromosomal VSG genes. This flanking sequence has three distinct sections: a 5' section containing 5 to 116 TAA or TAA-like tandem repeats; a moderately conserved sequence with an alternating GT characteristic; and a highly conserved 3' sequence that is largely alternating TA. Restriction mapping data and the location of this sequence suggest that it occurs at or near the 5' gene conversion boundary. This sequence, however, is not 5' to all VSG genes. These data suggest that this flanking sequence alone is not sufficient for VSG gene expression but that it may function in gene conversion for many, but not all, VSG genes.

INTRODUCTION

African trypanosomes contain numerous genes that encode immunologically distinct variant surface glycoproteins (VSGs). Most of these genes are located in tandem clusters within the chromosomes but some are adjacent to telomeres [1-3]. The VSGs are only produced during restricted stages of the trypanosome life cycle and only one VSG gene is transcribed in each trypanosome [see 4,5 for review]. The transcriptionally active VSG gene is invariably telomeric [5-7].

Trypanosomes periodically change antigenic type by spontaneously switching transcription to another VSG gene. This process, called antigenic variation, permits the trypanosome to evade the host immune response. The antigenic switch can be accomplished by either of two processes. Either a copy of a different VSG gene replaces the active VSG gene or transcription is switched to a VSG gene located on another telomere [8-13]. The first process has the characteristics of gene conversion and entails duplication of a basic copy (BC) of the gene, which may be either intrachromosomal or telomeric. Any one of several different telomeres may be used to express VSG genes [14-16].

The gene conversion is initiated between 1 and 5 kb 5' to the coding sequence of intrachromosomal BCs and an analogous location in the 5' flank of the converted sequence [17-20; Aline *et al.*, unpublished results]. This results in duplication of some 5'-flanking sequence of the BC gene which is called the cotransposed segment. The gene conversion terminates near the 3' terminus of the VSG gene coding sequence [see 5]. Telomeric BC genes can be duplicated in an analogous manner but some telomeric gene conversions appear to entail large DNA segments [21].

Although the exact 5' limit of the gene conversion is uncertain, several imperfect 76 bp repeats occur about 1.5 kb 5' to several BC VSG gene coding sequences [1,19,20]. Borst and his colleagues have suggested that these repeated sequences may function as recognition sites for gene duplication [19]. They have also suggested that other sequences occurring at the 3' end of VSG genes may have an analogous function [5]. Michiels *et al.* [18], Donelson *et al.* [22], and Pays *et al.* [23] have reported VSG gene duplications in which the 5' terminus does not contain 76 bp repeats; this suggests that either these gene conversions were initiated 3' to the 76 bp repeats or there are sequences other than 76 bp repeats on the 5' flanks of some VSG genes which serve an analogous function. Campbell *et al.* [20] have reported the presence of a (TAA)₉₀ sequence within one of the 76 bp repeats of an expressed VSG gene and suggested that this inserted sequence may play a role in VSG gene expression. In an effort to determine the function of 5'-flanking sequences we have examined the 5' flanks of several intrachromosomal VSG genes. Most, but not all, of these genes are flanked by the 76 bp repeat sequence. For several genes these sequences contain a long series of tandem TAA repeats that varies in size among the genes. These studies support the concept that the 76 bp repeat functions in gene conversion rather than VSG gene expression but also suggest that sequences other than the 76 bp repeat may have an analogous function for some VSG genes. These results also indicate that a TAA tandem repeat sequence is not sufficient for VSG gene expression.

METHODS

Trypanosomes: The production and characterization of the IsTaR 1 serodeme has been described in detail elsewhere [12,24]. By convention each variant antigenic type (VAT) is designated by a letter or number following a (.) in each serodeme (i.e., IsTat 1.A). For convenience only the letter or number following the (.) is used for IsTat VATs in this report. The superscript in VAT A⁵ designates the previous VAT (VAT 5).

The isolation and characterization of the A, 3, and 5 VSG gene cDNAs and the preliminary characterization of VSG gene families are described in Parsons *et al.* [25]. A capital letter following a (-) is used to distinguish members of the same VSG gene family (i.e., 5-B).

Southern Hybridizations: Trypanosome genomic DNA was prepared as described in Milhausen *et al.* [26]. Restriction enzyme digestions were performed according to supplier's instructions (Bethesda Research Laboratories, Inc.). Agarose gel fractionation of the DNA, transfer to nitrocellulose, and hybridization to nick-translated cDNAs were performed as described in Myler *et al.* [14] except where noted. The 76 bp repeat probe used here was prepared as a *Sau* 3A-Hinf I fragment from a plasmid containing the 5' flank of the expressed MITat 117 VSG gene [20].

Preparation of Genomic Libraries: Trypanosome genomic DNA was partially digested with varying concentrations of *Bam* HI or *Mbo* I. The resulting digests were mixed, size fractionated on preparative agarose gels, and 11 to 25 kb DNA fragments isolated by electroelution. This DNA was then ligated into *Bam* HI digested lambda 1059 DNA, packaged *in vitro*, and used to infect *E. coli* strain Q359 as described in Karn *et al.* [27]. The resulting phage were amplified on strain Q358 to produce high-titer libraries containing approximately 10^8 recombinants per ml. Genomic clones containing A, 3, and 5 VSG family member sequences were identified by the plaque hybridization method of Benton and Davis [28] using A, 3, and 5 cDNA clones as probes.

Subcloning: DNA from recombinant lambda 1059 phage containing A, 3, or 5 VSG gene family members was prepared as described in Maniatis *et al.* [29]. Positions of the VSG gene coding sequences were determined by hybridization of the cDNAs to restriction digested phage DNA. Restriction fragments containing 5'- or 3'-flanking sequences were isolated by electroelution, ligated into appropriately digested pBR322 DNA, and used to transform *E. coli* strain HB101 to ampicillin resistance. Recombinants were screened by restriction enzyme digests of small-scale plasmid DNA preparations. Large scale preparation of plasmid DNA was performed using the alkaline lysis procedure described in Maniatis *et al.* [29].

DNA Sequence Analysis: Nucleotide sequences of cloned fragments were determined using the M13 chain termination method [30]. *Bal* 31 digestion, used to produce a series of fragments convenient for sequence analysis, was performed as described in Payne *et al.* [31].

RESULTS

Five separate genomic segments, each containing a VSG gene family member,

Figure 1: Restriction maps of intrachromosomal VSG genes cloned into lambda 1059. The location of the coding sequence for each gene that is shown by the solid bar beneath the map was determined by Southern hybridization using VSG cDNAs as probes. Open boxes indicate hybridization to the pBS35 probe; solid boxes show hybridization to the 76 bp probe. Hybridizations using the 76 bp probe were washed at 37°. Restriction endonuclease abbreviations are: B, Bam HI; Bg, Bgl I; B1, Bgl II; C, Cla I; E, Eco RI; Ec, Eco RV; H, Hind III; K, Kpn I; Mb, Mbo I; P, Pst I; Pu, Pvu II; Pv, Pvu I; S, Sal I; Sm, Sma I; Sp, Sph I; Ss, Sst I.

were isolated as overlapping clones from the lambda 1059 libraries. One segment cloned from variant antigenic type (VAT) A DNA contains the single intrachromosomal A gene, labelled A-B, which differs from the A cDNA and the two telomeric A genes which can be expressed [12]. Two segments cloned from VAT 3 DNA contain two of the approximately 10 intrachromosomal 3 VSG gene family members [32] and are labelled 3-B and 3-C. Two segments cloned from VAT 5 DNA contain two of the three intrachromosomal 5 VSG gene family members [25] and are labelled 5-B and 5-C. The restriction maps and the VSG coding sequence locations of these five genes are shown in Fig. 1. The restriction maps of the coding sequences of these five genes differ from the corresponding cDNAs and expressed genes indicating that none of these genes are the expressed

Figure 2: Restriction maps and sequencing strategies of segments subcloned from the clones indicated in Fig. 1. Plasmids pBS35 and pBB46 were subcloned directly from the lambda 1059 clones while pBH14 was derived from a subclone of the larger Bam HI fragment shown. The location of the putative 110 VSG gene in pBS35 is suggested from sequence data. The arrows indicate the direction and limit of sequence obtained from Bal 31 digested plasmid DNA. Restriction endonuclease abbreviations are as in Fig. 1 and: AC, Acc I; Av, Ava I; Bt, Bst NI; Hc, Hinc II; Hd, Hind III; R, Rsa I.

telomeric copy (TC) or are BCs that were exactly duplicated to produce the TCs.

In order to study sequences which might be involved in the gene duplication mechanism of VSG gene activation, we examined the sequence relatedness of VSG gene flanking sequences. Subclones of VSG gene flanking sequences were used to probe Southern blots of the lambda 1059 clones. The Bam HI - Sal I fragment from the 5' flank of the 5-B gene, subcloned to produce plasmid pBS35, hybridized to four regions of the five cloned genomic segments (open boxes, Fig. 1). It hybridized strongly to the 5' flank of the 5-B gene (as expected), the 5' flank of the 3-B gene and the 3' flank of the A-B gene. It also hybridized, although less strongly, to the 3' flank of the 5-C gene. In order to avoid cross-hybridization due to homologies in the 5'

pBS35 (5-B):

```

<----110 VSG----
AGCCCTCTTTTGTGCTTGCCTTTTGTGTGCTGAATTAATTTACTAAAAAAGGGGTTTTTC -271
      CCTC
ACTTTTAAAAAATTTATAAAAAATTGATATTGAGAAATTTAAAAAATTATATAATATT -211
      APTm      8-mer
TGATTTTATTTTAACTTTTAACACAAGGTAAGCAAATTATAGAAACGTTGGGAATTTAA -151
      14-mer
TAATCTTTATATTTAAATCATGAAAATTTGTTATTATAGTGTTCGATATAAAGCAG -91
      ClaI
TATTATGACATATTATAAAAAAATAGTGTATTCTCAAACCTTTTCAATTTTGCTTATTC -31
TATTTATTTTATTATCAATAGGGTAATTCG

```

```

+1
CAG      TTATAATAGTAACAGTAATAG
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAG

```

```

AAGAGTGTTATGAGTGCATATATACGAATATTATGACGCGAGTGATGGTGAAGACTATAT 401
GGACTAATAGAATAGAAGCACTAAAAAATGAAGACAAGGCTTCAATCGGATAAGAAATGG 461
AGAGGGATAAAGATAGTACGACGTAAAAATTTAAAGGAAACGGAGATAAACGTGAAAAGA 521
      RsaI
GAAGGTGAAGGCATATGCACAGTTGTAAAAATGTGACGATGGCCATAATCATACACAGGA 581
AAATTATGTAAGGAAAAAECTCAAGGATTG

```

pBB46 (3-B):

GATATCGGTTAGGAACATAG
EcoRV

```

+1
CAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA

```

```

GAGA-T-TT-TG--T-TGTATATACTAATATTATAGTAAGTGGTGGTTGGGACCATATTG 216
TAGAAAAAATGTGAGCCTTTTAAATAGTAAAAACCC

```

pBH14 (A-B):

```

AGACAGAACATGAGAAAATTTGCAAAATAGGAAAGCGACGCTTTCAGGAAGCCAAGTTTCT -172
TTCCAAATAAGAAATTTGTCTGTTTTTCTGCTGCTTTTATCATTTTGGAAAGCTTTTAGT -112
      HdIII
ATTTCAAAACACAAAGCATTTTGTCTGATATTATGATAATTATGATAAAA -55
TAA      TGATATTGATAATAG
GAAAGTGCTGCGGGTGTGTATATACGAATATTGTAATAAGAG

```

```

+1
CAG          TAATAATAATAATGATAATAATAATAGTAATAG
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
TAATAATAATAAGATAATAATGATAATACTAATAATAATAATGATAATAATA
ATAAGA

GTGAGTGGTTGTGAGAGTGTATATATACGAATATTATAATAGGAATGTTTGAATAGACTAA 319
ATTTGCAAGATAGAATAAAAGACATTAAAACAATGGAAAGGATGAAATAAGAGTGATTAA 379
AAGAGTAGGATATTGAAAGTGGGAGAGCATGAGAAATATTATTAATAGTAACACAGAAAG 439
ACATAAGAGGAAGAAAGAGGGGGACGGAATATGCTCAGTCTTAAACAAAAGATGATAAT 499
GGTCGCTACTACACACAATGAAATGAGAGAGGAATAGCTTATAGAATTATGGCAAATTGG 559
TGGCACTACATCATCAGCAGAAGAAGCGGTAATAGGTCAACAATAAGAAGAAAAATATGT 619
AAAAAAACGACTGAATAGGCGCAGTGCAGAGCATATGTAAATAGAGCATCATGAGAAGA 679
TTGTGCTCCACATTTCAATCAAAAATAGATAGATGCAACTGTGAAGATGAGAGTGA 739
GATTTTAAATCTGATGGCGGATCC
          BamHI

```

Figure 3: Nucleotide sequences of the regions of interest in pBS35, pBB46, and pBH14. The CAG preceding the long series of tandem TAA repeats in each sequence is designated position 1. The 76 bp sequences are indicated by bold type. The characteristic 3' sequences (CCTC, A_nT_m , 8-mer, 14-mer) defining the 3' end of the putative 110 VSG gene are boxed. Important restriction sites are underlined.

ends of the 5-B and 5-C genes, a fragment of pBS35 lacking the 5' end of the 5-B gene was used to probe the 5-C gene. Thus, there are sequences on the 5' flank of the 5-B gene that are also present on the flanks of other VSG genes.

A probe containing five copies of the 76 bp sequence from the 5' flank of the expressed MITat 117 TC was used to determine whether or not the hybridization with pBS35 was due to this repeated sequence. This probe hybridized strongly with the 5' flank of the A-B gene and both the 5' and 3' flanks of the 3-C gene (closed boxes, Fig. 1). It also hybridized weakly with the 3' flank of the 5-B gene, but the site of hybridization was not localized and is therefore not shown. The 76 bp probe also hybridized very weakly with the same genomic segments which hybridized with pBS35. This probe also hybridized with a large number of bands in digests of trypanosome genomic DNA (Fig. 4A). Since the 76 bp probe hybridizes with eight different regions in the lambda 1059 clones and with numerous genomic fragments, it occurs frequently in the trypanosome genome.

In order to characterize the relationship among the genomic segments that

Figure 4: Hybridization of the 76 bp probe and a TAA probe to digests of trypanosome genomic DNA. The TAA probe is a Cla I - Rsa I restriction fragment from pBS35 containing 109 TAA repeats, one 76 bp sequence and approximately 200 nucleotides of other sequence (see Fig. 2). Panel A: 76 bp probe; lanes 1-3, Pst I digests of VAT 3, 5, and A⁵ DNAs respectively; lanes 4-6, Eco RI digests of the same DNAs; lanes 7-9, Bam HI digests of the same DNAs. Panel B: TAA probe; lanes as in panel A.

hybridize strongly to the pBS35 probe but weakly to the 76 bp probe, two of these segments were subcloned. These segments were mapped in greater detail, the region of cross-hybridization more accurately localized and that region sequenced, as shown in Fig. 2. Each nucleotide was determined from the sequence of at least two independently isolated M13 clones. The Bam HI - Hind III fragment 3' to the A-B gene and the 5' Bgl II fragment of the 3-B gene were subcloned to produce plasmids pBH14 and pBB46, respectively. The presence of a VSG gene, arbitrarily designated 110, 5' to the 5-B gene was inferred from sequences defining the 3' end of a VSG gene as discussed below. The nucleotide sequences of the regions of interest are shown in Fig. 3.

As expected, each of the sequenced regions contains the 76 bp sequence. In addition, each region also contains a long series of tandem TAA repeats. The number of repeats varies among the three sequences. There are 109 repeats in pBS35, 58 in pBH14, and 52 in pBB46. In each case the TAA repeats are inserted in the same relative position within the 5' end of the 76 bp sequence (shown in bold type, Fig. 3). There is a single 76 bp sequence in both pBS35 (5-B gene) and pBB46 (3-B gene) which indicates that a single copy of this sequence can be detected with the 76 bp repeat probe. There are two copies of the 76 bp sequence in pBH14 (A-B gene) but only the 3' copy contains a long series of TAA repeats whereas the 5' copy contains a 15 bp TAA-like sequence at that position.

While the 76 bp probe hybridizes to a large number of bands (Fig. 4A), the TAA probe, under stringency conditions which allow hybridization to extensive TAA repeats but not to 76 bp sequences, hybridizes to only about 30 bands in digests of trypanosome genomic DNA (Fig. 4B). Thus, extensive TAA repeats occur less frequently in trypanosome genomic fragments than do 76 bp sequences. Since the bands in Fig. 4B appear to be a subset of those in Fig. 4A and all trypanosome genomic segments cloned into lambda 1059 which hybridize to the TAA probe also hybridize to the 76 bp probe, it seems likely that all TAA repeats are associated with, and most likely inserted in, 76 bp sequences. Thus, 76 bp sequences can be up to at least 400 nucleotides in length when the TAA repeats are included, as is the case for the 5-B gene. In view of this variation in length of what have been called 76 bp repeats (and association with VSG genes) we shall refer to them as V-sequences for simplicity. By comparing our four V-sequences with the five V-sequences of the MITat 118 BC [19], the three V-sequences of the MITat 117 BC and the seventeen V-sequences of the expressed 117 TC [20], we generated the consensus V-sequence shown in Fig. 5. Since there are only single copies of the V-sequence on the flanks of the 5-B and 3-B genes, we are able to assign 5' and 3' termini to the V-sequence. On the basis of nucleotide composition we have divided the V-sequence into three sections. The 5' section, section I, is a TAA-like sequence consisting of a variable number of TPuPu repeats, from as few as 5 to as many as 116, usually beginning with a CAG and ending in a TAG. The long series of TAA repeats are always found here and are in the same relative position as the (TAA)₉₀ sequence on the 5' flank of the expressed MITat 117 TC [20]. The middle region, section II, is a 20 bp sequence which is predominately alternating GT and is moderately conserved. Section III, which constitutes the 3' end of the V-sequence, is a 24 bp sequence which is

5' CAG[TAA]₄₋₁₁₅TAG*GAGAGTGTGTGAGTGTGTG*TATATACGAATATTATAATAAGAG 3'

Section I:

C A G [T A A] ₄₋₁₁₅ T A G										

20									C	

1	29	4						28	5	A

8							27			T

			25				2	1	23	G

								1	O	

Section II:

G	A	G	A	G	T	G	T	T	G	T	G	A	G	T	G	T	G	T	G	
<hr/>																				
2										1 1		1 1								C
<hr/>																				
7	27	1	25	2	1		1	1	3	3		26	2		2		1		A	
<hr/>																				
1		1	3	1	29	1	26	28	27		1	1	27		27		7		T	
<hr/>																				
22	1	27	1	25	27		26		1	23	1	28	29		25		7		G	
<hr/>																				
1				1	1								1	1	1		1	22	2	O

Section III:

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

1				29				1								1				1					C	

25		2	29		29		2	28	29	29				27	27		28	25		26	1	27	4	A		

26	26		29		5				28				29	29	1	29	28						1	1	T	

1				22				1						1		2		1	4		1	26	24		G	

3	3																				1		1	1	1	O

Figure 5: A consensus for the V-sequence determined from 29 V-sequences. The consensus sequence is divided into three regions by *s and the frequency of occurrence of each nucleotide is indicated below each position. O = a nucleotide deletion at that position. The # indicates a CT insertion at that position in one sequence. Included are 5 repeats from the MITat 118 BC [19], 3 from the MITat 117 BC [20], 17 from the MITat 117 expressed TC [20], and the 4 reported in this work.

predominately alternating TA and is highly conserved.

Since VSG genes occur in tandem arrays in the trypanosome genome [1], the 5' terminus of one VSG gene is adjacent to the 3' terminus of another. Thus, the V-sequences on the 3' flank of the A-B, 3-C, 5-B and 5-C genes (see

Figure 6: Hybridization of a 5'-flanking probe from the 5-B gene to digests of trypanosome genomic DNA. The probe is the 3' Bst NI - Rsa I restriction fragment from pBS35 (see Fig. 2). Lanes 1, 2, and 3: VAT 1 DNA digested with Hind III, Bgl I and Hind III, or Bgl II and Hind III, respectively. Lanes 4, 5, and 6: VAT 5 DNA digested as in lanes 1-3. The expressed 5' TC is denoted by a star.

Fig. 1) may represent 5'-flanking sequences of unidentified VSG genes. This possibility is supported by the occurrence of V-sequences on the 5' flank of the MITat 117 TC [20], a telomeric VSG gene with no VSG gene immediately 5' to it, indicating that the V-sequence is a 5' characteristic. In addition, there are sequences upstream of the 5-B gene V-sequence that are characteristic of the 3' end of VSG genes [5,6,19]. The CCTC, A_nT_m , 8-mer, and 14-mer sequences, conserved sequences that characterize the 3' end of VSG genes, present in the pBS35 sequence between nucleotides -326 to -197 (Fig. 3) define the 3' end of the putative 110 VSG gene. Thus, the V-sequence on the 5' flank of the 5-B gene is also 3' to an upstream VSG gene. These results confirm the clustering of VSG genes and support the suggestion that V-sequences 3' to VSG genes are 5'-flanking elements of adjacent VSG genes.

Figure 7: Restriction maps of the expressed 5 TC, the 5 BC, and the 5-B genes. The indicates additional adjacent DNA and the solid circle a chromosome end (telomere). Restriction endonuclease abbreviations are as in Figs. 1 and 2.

The 5'-flanking sequences of the 5-B gene, at least up to the first *Rsa* I site, are also present in the 5'-flanking sequences of the 5 BC gene and the expressed 5 TC. Using the segment from the *Bst* NI to the *Rsa* I sites of pBS35 as a probe, it was found that this region is conserved in all members of the 5 VSG gene family, except for the 5-C gene (Fig. 6). This can be clearly seen in the *Hind* III + *Bgl* II digests (lanes 3 and 6). The 5 BC gene is located on a minichromosome [13] and corresponds to the >50 kb band. The 9.4 kb band corresponds to the 5-B gene and the 2.9 kb band to the 5-D gene, the third intrachromosomal 5 VSG gene family member. The 22 kb band, present only in digests of VAT 5 DNA, corresponds to the expressed 5 TC. There is no hybridization to the 5-C gene in any of these or 7 other restriction endonuclease digests (data not shown). The absence of this conserved region in the flank of the 5-C gene explains why pBS35 failed to hybridize to the 5' flank of the 5-C gene (see Fig. 1). The conservation of this 5'-flanking sequence in the other members of the 5 gene family, however, allows us to estimate the 5' limit of gene conversions for these genes. As seen in Fig. 7, the 5' restriction maps of the 5 TC, the 5 BC, and the 5-B genes diverge beyond the *Rsa* I site. Therefore, gene conversion must have been initiated between this *Rsa* I site and the *Hinc* II site of the BC. The V-sequence occurs in an analogous position in the 5-B gene, between the *Rsa* I and *Cla* I sites (Fig. 2).

DISCUSSION

Expression of intrachromosomal VSG genes requires a gene conversion event resulting in insertion of a VSG gene copy into a telomeric expression site. We have examined the 5'-flanking sequences of several non-expressed intrachromosomal VSG genes. Although we do not know whether or not these genes are ever expressed, they provide an opportunity to identify and characterize sequences that may be essential to gene conversion or to expression of VSG genes. We have found that many, but not all, intrachromosomal VSG genes contain at least one copy of a sequence, which we call the V-sequence, 5' to the coding sequence. V-sequences have several notable characteristics. They occur between 1 and 1.5 kb 5' to VSG genes, a location that is at or near the 5' boundary of VSG gene conversion. The V-sequences can occur singly or in up to at least seventeen direct tandem repeats. The V-sequence has a tripartite structure. The length of the 5' section varies among repeats and can contain a striking run of tandem TAA repeats. Only one of the V-sequence repeats in a unit contains such a run.

The location of V-sequences at or near the 5' gene conversion terminus has led to the suggestion that they may serve as recognition sequences for a site-specific recombination [19,20]. Our data support this hypothesis. Although none of the genes we have examined are BCs of the expressed IsTaR 1 VSG genes that we have studied, the 5' flanks of the A-B and 5-B genes have sequence homology to the 5' flanks of the A and 5 genes which are expressed in several A VATs and two 5 VATs, respectively. Since the A-B and 5-B genes contain V-sequences in their 5' flanks, this implies that the expressed A and 5 genes, which are known to be expressed via the gene conversion mechanism, also contain V-sequences in their 5' flanks. In addition, by inference from restriction mapping data the V-sequence on the flank of the 5 BC gene occurs at or near the 5' terminus of VSG gene conversion in VAT 5.

The apparent absence of V-sequences 5' to some VSG genes, such as the 5-C gene, suggests that either these genes are incapable of gene conversion (and hence expression) or other sequences may function as 5' recognition sequences. Although we cannot rule out the possibility that the V-sequences, due to their unusual properties, were lost during cloning, there are two general possibilities that would permit gene conversion (and expression) of a VSG gene lacking V-sequences. A V-sequence may occur 5' to a VSG gene that is itself 5' to the VSG gene lacking the V-sequence. In this case a segment containing multiple VSG genes would be duplicated to a telomeric site. Such a situation could explain the large 5 kb cotransposed segment of the IsTaT 11 VSG gene

(Aline *et al.*, unpublished results) and would predict the presence of a VSG gene between the expressed telomeric gene and the 5' barren region. The second possibility is that sequences other than the V-sequence may serve as 5' recognition sequences. Such sequences might be related to the V-sequence but be sufficiently divergent that they would not hybridize with the V-sequence probes or they may be unrelated. The AnTat 1.1 [18,23] and the IaTat 1.3 [22] 5' duplication boundaries have been sequenced and are related to each other but not to the V-sequence. While this suggests that sequences unrelated to the V-sequence may serve as 5' gene conversion recognition sequences, this observation is complicated by the fact that the BCs of these genes are telomeric. This leaves open the possibility that these VSG genes may have been activated by the partial gene conversion mechanism first described by Pays *et al.* [33] and that V-sequences occur further 5'.

The significance of the sequence variation among V-sequences, including the dramatic variation in the length of the $(TAA)_n$ sequence and in the number of V-sequences associated with individual VSG genes, is unknown. Campbell *et al.* [20] have suggested that the $(TAA)_{90}$ sequence associated with the expressed MITat 117 TC might function as a transcription reinitiation site that would mediate the expression of this gene. However, the presence of up to 109 TAA repeats at the same position within the V-sequence flanking non-expressed VSG genes of the IsTaR 1 serodeme shows that this element alone or as part of a single V-sequence is insufficient for VSG gene expression. Although we cannot rule out the possibility that a long series of TAA repeats is necessary for VSG gene expression in conjunction with other control sequences present near expressed VSG genes, we find this possibility unlikely since the 1 VSG gene (IsTat 1.1) is expressed from a minichromosome which lacks a long series of TAA repeats [13]. Thus, we conclude that these TAA repeats are not independent sequence elements but rather constitute a subset of 76 bp sequences whose function remains unknown.

It is intriguing that the expressed MITat 117 TC is associated with 17 copies of the V-sequence, one of which contains a $(TAA)_{90}$ sequence, whereas intrachromosomal BCs either have several V-sequences with no long $(TAA)_n$ sequence (the MITat 117 and 118 BCs) or have only one or two V-sequences of which one contains a long $(TAA)_n$ sequence (the 3-B, 5-B, and 3' of the A-B gene). Presuming that the V-sequence is a gene conversion recognition sequence, it may be that the number of V-sequences associated with a VSG gene or the presence of a large $(TAA)_n$ sequence serves as an extensive homology block making the telomeric expression site a high frequency target for gene

conversion. On the other hand the variation in the number of V-sequences or length of the $(TAA)_n$ sequence may influence either the targeting of a VSG gene to particular telomeres or the probability that a VSG gene will gene convert a telomere. These possibilities alone or in combination could contribute to the quasi-sequential expression of VSG genes in a serodeme.

An alternative explanation for V-sequence variation is that unequal crossovers occur resulting in deletion or duplication of intervening V-sequences. Smith [34] has proposed that short tandemly repeated sequences promote unequal crossovers and such a mechanism has been proposed to explain the variation in the numbers of 36 bp repeat sequences present in yeast telomeres [35]. It is possible, therefore, that the variation in V-sequences merely reflects the continual occurrence of unequal crossover events.

Although the mechanism for VSG gene conversion is still unknown, the tripartite structure and particular sequence characteristics of the V-sequence, especially the striking TAA repeats, may provide some clues. The V-sequence might give rise to unusual physical characteristics of the DNA which may extend into adjacent DNA, and may even impart unusual characteristics to the chromatin containing this sequence. The potential for B to D conformational transitions in V-sequence DNA has been discussed in Liu *et al.* [19]. Perhaps more significant is the potential for B to Z conformational transitions in the TGTGTGTATATACG sequence at the junction between sections II and III of the V-sequence. Wang *et al.* [36] and Nordheim *et al.* [37] have shown that a short series of tandem pyrimidine-purine repeats can result in a Z DNA structure. Such a Z DNA structure in V-sequence DNA may serve as a recognition site for an endonuclease whose action initiates a gene conversion event in a manner analogous to yeast mating type switching [38]. The conservation of the TGTGTG sequence, which is found in eukaryotic transposable elements [39], may also have significance as noted in Campbell *et al.* [20]. The ability to mobilize intrachromosomal VSG genes for gene conversion (duplicative transposition) to telomeric expression sites and possibly the ability to target those genes to certain telomeres with varying probabilities has obvious adaptive value to the trypanosome.

ACKNOWLEDGEMENTS

We would like to thank Elke Gobright and Joanne West for excellent technical assistance and Dr. R. Nelson for help in the initial phases of this study. We would also like to thank Drs. N. Agabian and M. Parsons for helpful discussions and Dr. J. Feagin for critical reading of the manuscript. This research received support from NIH grant AI 17375, the World Health Organization, and the Murdoch Charitable Trust.

Abbreviations: BC - basic copy; bp - base pair; cDNA - complementary DNA; IsTat(R) - Issaquah Trypanozoon antigenic type (repertoire); kb - kilobase; Pu - purine; TC - telomeric copy (whether expressed or not); VAT - variant antigenic type; VSG - variant surface glycoprotein.

REFERENCES

1. Van der Ploeg, L.H.T., Valerio, D., De Lange, T., Bernards, A., Borst, P., and Grosveld, F.G.C. (1982) *Nucl. Acids Res.* 10, 5905-5923.
2. Williams, R.O., Young, J.R., and Majiwa, P.A.O. (1982) *Nature* 299, 417-421.
3. De Lange, T. and Borst, P. (1982) *Nature* 299, 451-453.
4. Parsons, M., Nelson, R.G., and Agabian, N. (1984) *Immunol. Today* 5, 43-50.
5. Borst, P. and Cross, G.A.M. (1982) *Cell* 29, 291-303.
6. Borst, P. (1983) in *Mobile Genetic Elements* (Shapiro, J.A., ed.) pp 621-659. Academic Press, New York.
7. Borst, P., Bernards, A., Van der Ploeg, L.H.T., Michels, P.A.M., Liu, A.Y.C., De Lange, T., Sloof, P., Veeneman, G.H., Tromp, M.C., and Van Boom, J.H. (1983) in *Genetic Rearrangement* (Chater, K.F., Cullis, C.A., Hopwood, D.A., Johnston, A.A.W.B., Woolhouse, H.W., eds.) pp 207-233. Croom-Helm, London.
8. Hoeijmakers, J.H.J., Frasch, A.C.C., Bernards, A., Borst, P. and Cross, G.A.M. (1980) *Nature* 284, 78-80.
9. Pays, E., Van Meirvenne, N., Le Ray, D., and Steinert, M. (1981) *Proc. Natl. Acad. Sci. USA* 78, 2673-2677.
10. Longacre, S., Hibner, U., Raibaud, A., Eisen, H., Baltz, T., Giroud, C., and Baltz, D. (1983) *Molec. Cell. Biol.* 3, 399-409.
11. Young, J.R., Donelson, J.E., Majiwa, P.A.O., Shapiro, S.Z., and Williams, R.O. (1982) *Nucl. Acids Res.* 10, 803-819.
12. Myler, P., Nelson, R.G., Agabian, N., and Stuart, K. (1984) *Nature* 309, 282-284.
13. Rothwell, V., Aline, R. Jr., Parsons, M., Agabian, N., and Stuart, K. (1985) *Nature*, 313, 595-597.
14. Myler, P., Allison, J., Agabian, N., and Stuart, K. (1984) *Cell* 39, 203-211.
15. Michels, P.A.M., Van der Ploeg, L.H.T., Liu, A.Y.C., and Borst, P. (1984) *EMBO J.* 3, 1345-1351.
16. Van der Ploeg, L.H.T., Schwartz, D.C., Cantor, C.R., and Borst, P. (1984) *Cell* 37, 77-84.
17. Boothroyd, J.C. and Cross, G.A.M. (1982) *Gene* 20, 281-289.
18. Michiels, F., Matthyssens, G., Kronenberger, P., Pays, E., Dero, B., Van Assel, S., Darville, M., Cravador, M., Steinert, M., and Hamers, R. (1983) *EMBO J.* 2, 1185-1192.
19. Liu, A.Y.C., Van der Ploeg, L.H.T., Rijsewijk, F.A.M., and Borst, P. (1983) *J. Mol. Biol.* 167, 57-75.
20. Campbell, D.A., Van Bree, M.P., and Boothroyd, J.C. (1984) *Nucl. Acids Res.* 12, 2759-2774.
21. Van der Ploeg, L.H.T., Cornelissen, A.W.C.A., Michels, P.A.M., and Borst, P. (1984) *Cell* 39, 213-221.
22. Donelson, J.E., Murphy, W.J., Brentano, S.T., Rice-Ficht, A.C., and Cain, G.D. (1984) *J. Cell. Biochem.* 23, 1-12.
23. Pays, E., Van Assel, S., Laurent, M., Dero, B., Michiels, F., Kronenberger, P., Matthyssens, G., Van Meirvenne, N., Le Ray, D., and Steinert, M. (1983) *Cell* 34, 359-369.
24. Stuart, K., Gobright, E., Jenni, L., Milhausen, M., Thomashow, L., and Agabian, N. (1984) *J. Parasitol.* 70, 747-754.

25. Parsons, M., Nelson, R.G., Newport, G., Milhausen, M., Stuart, K., and Agabian, N. (1983) *Mol. Biochem. Parasitol.* 9, 255-269.
26. Milhausen, M., Nelson, R.G., Parsons, M., Newport, G., Stuart, K., and Agabian, N. (1983) *Mol. Biochem. Parasitol.* 9, 241-254.
27. Karn, J., Brenner, S., Barnett, L., and Cesareni, G. (1980) *Proc. Natl. Acad. Sci. USA* 77, 5172-5176.
28. Benton, W.D. and Davis, R.W. (1977) *Science* 196, 180-182.
29. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, New York.
30. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
31. Payne, M., Rothwell, V., Jasmer, D.P., Feagin, J.E., and Stuart, K. (1985) *Mol. Biochem. Parasitol.*, in press.
32. Allison, J., Rothwell, V., Newport, G., Agabian, N., and Stuart, K. (1984) *Nucl. Acids Res.* 12, 9051-9066.
33. Pays, E., Van Assel, S., Laurent, M., Darville, M., Vervoort, T., Van Meirvenne, N., and Steinert, M. (1983) *Cell* 34, 371-381.
34. Smith, G.P. (1976) *Science* 191, 528-535.
35. Horowitz, H. and Haber, J.E. (1984) *Nucl. Acids Res.* 12, 7105-7122.
36. Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., Van Boom, J.H., Van der Mahel, G., and Rich, A. (1979) *Nature* 282, 680-686.
37. Nordheim, A., Lafer, E.M., Peck, L.J., Wang, J.C., Stollar, B.D., and Rich, A. (1982) *Cell* 31, 309-318.
38. Strathern, J.N., Klar, A.J.S., Hicks, J.B., Abraham, J.A., Ivy, J.M., Nasymth, K.A., and McGill, C. (1982) *Cell* 31, 183-192.
39. Varmus, H.E. (1983) in *Mobile Genetic Elements* (Shapiro, J.A., ed.) pp 411-503. Academic Press, New York.