Volume 12 Number 6 1984 Nucleic Acids Research

The 5'-limit of transposition and upstream barren region of a trypanosome VSG gene: tandem 76 base-pair repeats flanking $(TAA)_{90}$

David A.Campbell, Mark P.van Bree and John C.Boothroyd

Department of Medical Microbiology, Stanford University School of Medicine, Stanford, CA 94305, USA

Received 12 December 1983; Revised and Accepted 17 February 1984

ABSTRACT

We have cloned and sequenced a portion of the region upstream of an expressed VSG gene of Trypanosoma brucei. This "expression-linked copy" arose through the duplication and transposition of a silent, "basic copy" of the gene to an expression site. Comparison of the sequences surrounding the 5'-end of the transposed segment in the two loci indicates the 5'-limit of transposition lies within the first (3'-most) of three repeated segments found at this position in the basic copy locus. These highly conserved repeat segments which average 76 base-pairs in length are also found tandemly repeated upstream of the transposed segment in the expression site. In this latter site, however, they are more numerous (at least 17 repeats) and they are interrupted, within the middle of one repeat, by a 270 base-pair region consisting of $(TAA)_{90}$. The possible roles of these unusual sequences in transposition and in a model proposing discontinuous transcription of VSG genes are discussed.

INTRODUCTION

Antigenic variation in Trypanosoma brucei describes the ability of the organism, in successive waves of parasitemia, to change the species of glycoprotein which constitutes its outer coat. This variation is explained by the quasi-sequential expression of different variant surface glycoproteins (VSG's) from a repertoire of 10^2-10^3 VSG genes (for review, see Ref. 1-3). Activation of VSG genes fits a two-step model whereby a silent, basic copy (BC) of the gene is first duplicated and transposed to one of a few expression sites, thus forming the expression-linked copy (ELC), and then that ELC is activated by an as yet unknown mechanism. According to this model, the so-called "non-duplication activated" genes (1) have already achieved the first stage (i.e., they have already produced an ELC resident in an expression site) and now await the activation event. Occasionally, such genes have also lost their corresponding BC (1). The transposed segment (TS) is discrete in size, and in addition to the structural VSG gene typically includes a cryptic region of about 1.5 kb

upstream of the gene (4-6). In some instances, this additional segment possesses an open reading frame (4,5) but this has not yet been shown to have a coding function *in vivo*.

At least three expression sites appear to exist within the trypanosome genome (7; where an expression site is defined as a site with the *potential* for expression). These are characterized by two unusual properties: first, they are located within 5-10 kilobase-pairs (Kb) of a double-stranded break in the DNA (sometimes referred to as a chromosome-end or telomere; 8,9) and, secondly, the region between the gene and this break is essentially devoid of restriction sites as is a similar-sized region immediately upstream of the transposed segment (10,11). These constitute the so-called barren regions. Each of the expression sites, it appears, may simultaneously contain an intact VSG gene. Yet, as only one VSG species is found on the surface of a trypanosome at any one time, activation of one of these expression sites seemingly precludes activation of the others.

The way in which a given expression site is activated is not known, although some clues do exist. The mRNAs derived from different expression sites possess the same 35-nucleotides at their 5'-ends (4). The DNA coding for this leader segment (the so-called mini-exon) is not contiguously arranged with the protein-coding portion of the gene and surprisingly, no physical linkage between the two has yet been demonstrated (6, 12, 13). Instead, the mini-exon is found within a 1.35 kb unit which is tandemly repeated 100-200 fold. These 1.35 kb repeats are highly conserved and possess the consensus sequences found for eukaryotic promoters and, therefore, seem likely to be sites of transcriptional initiation (12). The process by which transcription initiated at these promoters culminates in an mRNA coding for a VSG remains an enigma.

To better understand the mechanism of transposition and in the hope of finding some further clues to the basis of expression site-activation, we have cloned and sequenced the 5'-limit of transposition and a portion of the upstream barren region for an expressed VSG gene.

MATERIALS AND METHODS

Trypanosome Strains and Preparation of Trypanosome DNA

Three cloned variants were used in this work: MITat (Molteno Institute Trypanozoon antigenic type) 1.4, MITat 1.5 and MITat 1.2 which express VSGs 117, 118 and 221, respectively. The generation of these variants (which for simplicity will be referred to by their VSG type) from a cloned isolate has

been previously described (14). Bloodstream forms of the parasites were prepared from infected rat blood and further purified as described (14).

Trypanosome DNA was prepared either by lysis of the cells in SDS (sodium dodecyl sulphate), digestion with proteinase K and ribonuclease, followed by phenol extraction and ethanol precipitation (15) or by cesium chloride gradient centrifugation of SDS-lysed cells treated with proteinase K and ribonuclease, essentially as described by Simpson, et al. (16). The results obtained with DNA prepared by the two methods were indistinguishable.

Genomic Mapping

The conditions for restriction endonuclease digestions were as specified by the suppliers (New England Biolabs). Digested DNA was electrophoresed in 0.7% (w/v) horizontal agarose gels (17) containing Tris-Borate-EDTA (pH 8.3) buffer and ethidium bromide. Pst1-digested bacteriophage λ DNA was used for size markers. Size-fractionated DNA was transferred to nitrocellulose according to the method of Southern (18). Specific DNA probes were isolated by preparative agarose gel electrophoresis followed by electroelution in 1/20 Tris-Acetate-EDTA buffer (17) and then labelled with ^{32}P by nick-translation (19). Hybridization of filters with probe was performed at 65°C in 3 x SSC (1 x SSC = 0.15M sodium chloride, 0.015M sodium citrate, pH 7.0; 18). Washing was at various stringencies as detailed in the figure legends.

Molecular Cloning and Plasmid Purification

DNA to be used in cloning was digested with the appropriate restriction endonuclease and resolved by preparative agarose gel electrophesis. Fragments of the required size were excised and purified as described above. These were ligated into the appropriately digested vector and used to transform E. coli HB101 to ampicillin resistance (20). Two vectors were used in this work, pAT153 (21) and pUC8 (22). Screening of recombinants was initially by colony hybridization (23) using DNA fragments labelled with ^{32}P by nick-translation (19). Subsequent screening was by restriction enzyme analysis of small-scale preparations of plasmid (24). Large scale preparations of recombinant plasmids thus identified were made by two isopycnic CsCl gradient centrifugations of SDS-lysed bacterial lysates as described (25).

Deletion derivatives of selected recombinants were generated to facilitate DNA sequence analysis. These were produced by linearisation of the plasmid at a unique site followed by digestion with nuclease Bal31 for a

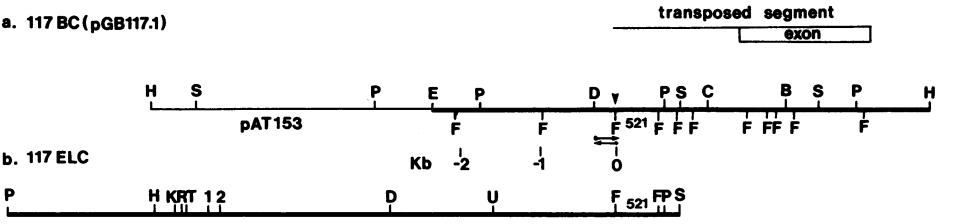

Figure 1. Restriction maps of 117BC and 117ELC. a. Restriction map of the recombinant plasmid pGB117.1 containing the 117BC. The generation of this recombinant using the plasmid vector pAT153 has been previously described (4). All sites within the insert for the following enzymes have been shown: B, BglI; C, ClaI; E, EcoRI; F, HinfI; H, HindIII; P, PstI; S, SalI. The critical DdeI site at position -289 (see text) is also shown. Distances are in kilobase-pairs (Kb) from the indicated HinfI site (▼). The regions where useful sequence information was obtained from this plasmid is shown by dotted arrows (the filled-in dots represent the 5'-^{32}P end label and the arrows indicate the extent of the region where unambiguous sequence was obtained). '521' indicates a HinfI fragment used for hybridisation in subsequent experiments. The regions corresponding to the 117 transposed segment and VSG-coding exon are also shown. Thin and thick lines represent plasmid vector and trypanosome-derived DNA, respectively. b. Restriciton map of the region upstream of the 117ELC as deduced by Southern blot analysis using the HinfI-521 fragment from pGB117.1 as the probe. Abbreviations are as above with additional sites: U, Sau3A; 1, HhaI; 2, HincII; T, TaqI; R, RsaI; and K, KpnI. Only the first site upstream of position 0 could be determined for each enzyme by this method.

limited time. The resulting material was size-fractionated on agarose gels and recircularized by ligation in the presence of EcoRI linkers. Transformation and screening were as above.

DNA Sequence Analysis

Nucleotide sequences were determined according to the method of Maxam and Gilbert (26) using 5'-^{32}P-labelled restriction fragments. Bal 31 deleted clones (see above) were used to sequence through regions where there were no convenient restriction sites. Five sequencing reactions (G, A+G, A>C, C+T and C) were employed and the 0.4 mm gels (27) were usually fixed in 10% (v/v) acetic acid and dried prior to autoradioagraphy (28).

RESULTS

Mapping of the 5'-limit of Transposition

To determine the approximate location of the 5'-limit of transposition, it was necessary to generate detailed restriction maps of the loci containing the 117 BC and 117 ELC in the relevant regions. We already possessed such a map for a cloned copy of the 117 BC (pGB117.1; ref.4) and

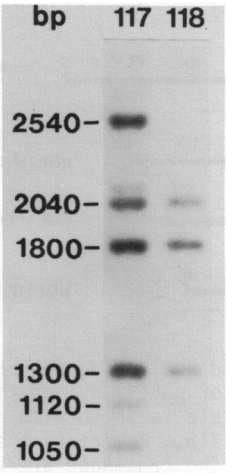

Figure 2. Identification of a restriction fragment suitable for cloning the region upstream of the 117ELC. About 1.5 µg of genomic DNA from variants 117 and 118 was digested with Sal I and Sau3A and resolved by electrophoresis in a 0.7% agarose gel. The DNA was transferred to nitrocellulose and hybridised to 32P-labelled HinfI-521 fragment from pGB117.1 (see fig. 1a). Washing was in 0.1 X SSC at 50°C. Sizes are in base-pairs (bp) by comparison to a PstI digest of bacteriophage λ DNA (not shown). The 117 BC and 117 ELC are estimated at 2040 and 2540 bp, respectively.

expanded this to include the critical DdeI site (fig. 1a). No clone of the complete 117 ELC existed despite attempts by ourselves and others to produce one. To generate a map, therefore, we used "Southern" blots (18) of genomic DNA digested with several restriction enzymes and probed these with the HinfI-521 fragment from pGB117.1 (data not shown). This probe represents the 5'-end of the 117 transposed segment and so by comparing 117 DNA, possessing the 117 ELC, with 118 DNA which lacks the 117 ELC, we were able to generate the map shown in figure 1b.

This map served two purposes. Firstly, it localized the 5'-limit of transposition to between the DdeI site at position -289 in the 117 BC clone (which is not observed in the 117 ELC map) and the HinfI site at position +1 (which is found in both the BC and ELC maps). Second, it revealed a Sau3A site within the barren region which, on the assumption that failure to clone the complete ELC was due to the deleterious effect of the complete barren region on a plasmid replicon, might enable a portion of the barren region and the 5'-limit of transposition to be cloned. Further mapping revealed that a Sau3A/SalI double digest of 117 DNA would resolve the 117 ELC well

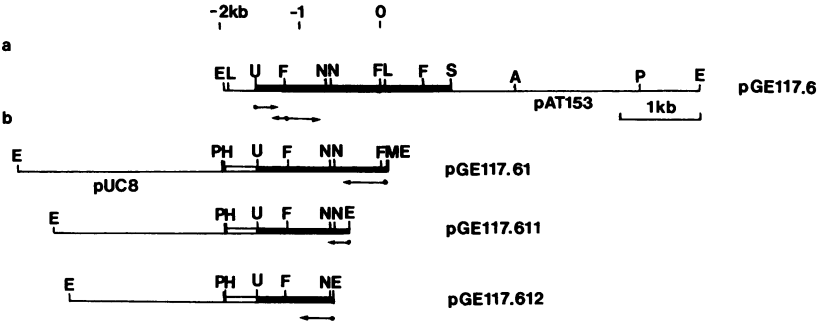

Figure 3. Restriction maps and nucleotide sequencing strategy for 117ELC plasmids. a. Restriction map of pGE117.6 linearised at the EcoRI site of the plasmid vector pAT153. Scale is in Kb with numbering from the HinfI site near the 5'-end of the 117TS. Restriction site abbreviations are M, BamHI; N, Fnu4H; and L, AluI; the remainder are as for figure 1. All sites for these enzymes within the insert and upstream of position 0 are shown. The two AluI sites used for subcloning (see text) are indicated. b. Nucleotide sequencing strategy used for the remainder of the cloned barren region. The generation of the two subclones, pGE117.611 and pGE117.612 are described in the text. Dotted arrows represent the regions where useful sequence information was obtained from these plasmids. The open boxed region represents vector-derived material from pGE117.6 carried across in the subcloning.

clear of the 117 BC and other 117 family members in a fragment of about 2.5 Kb (figure 2).

Molecular Cloning of the 5'-end of of the 117 ELC

The Sau3A/SalI fragment of about 2.5 kb described above was prepared by preparative agarose gel electrophoresis and ligated into the plasmid vector pAT153 which had been digested with BamHI (an accepting site for Sau3A fragments) and SalI. The resulting recombinants were used to transform E. coli HB101. 117 ELC-containing clones were selected by colony hybridization using the HinfI-521 fragment of pGB117.1 as a radio-labelled probe. Four recombinants gave positive signals in two separate attempts, and of these, two were chosen for further characterization (one from each attempt). These two recombinants, which will be referred to as pGE117.6 and pGE117.9, were indistinguishable by extensive restriction enzyme analysis (data not shown). pGE117.6 was chosen for further study and its restriction map is shown in figure 3.

Precise Localization of 5'-limit of Transposition

To identify the precise 5'-limit of transposition, it was necessary to determine the complete nucleotide sequence for both the 117 ELC and 117 BC

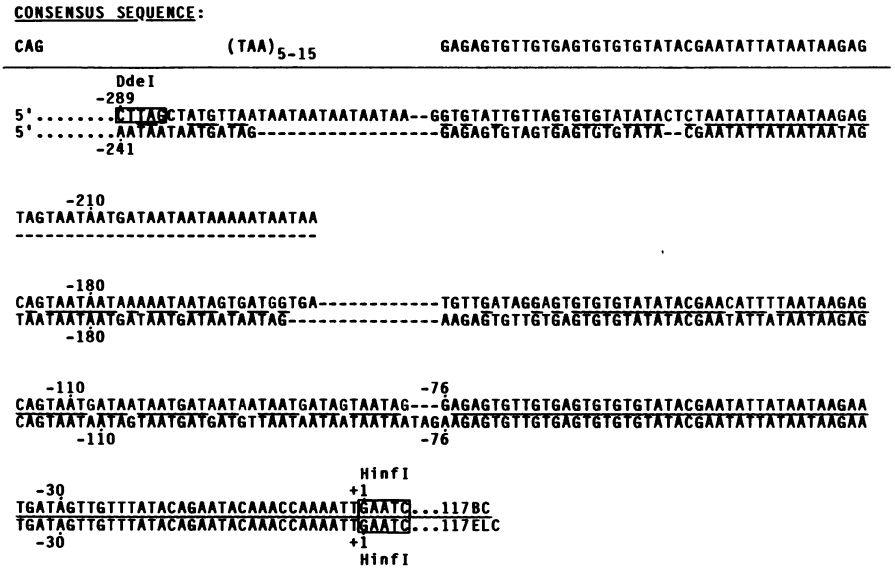

Figure 4. Nucleotide sequence in the region containing the 5'-limit of transposition. The sequences obtained for the regions upstream of the HinfI site designated position 1 are presented for the 117BC- and 117ELC clones on the top and bottom lines, respectively. Homology between the sequences is indicated by a line between them. Gaps have been introduced to maximize the homology and the critical restriction sites are boxed and labelled. The sequences are arranged to convey the 76-bp repeat motif and gaps are again introduced to illustrate the basic pattern of this motif as presented in the consensus sequence at the top of the figure.

within the previously determined, approximate limits defined above. This was done using the chemical modification method of Maxam and Gilbert (26) according to the strategies shown in figures 1 and 3 for the 117 BC and 117 ELC, respectively. The results are presented in figure 4 and show that the 5'-limit of transposition occurs within the first (3'-most) of three highly conserved repeats found between the DdeI and HinfI sites of the 117 BC, located at positions -289 and +1, respectively. The exact nucleotide where the cross-over occurs cannot be identified because of the high homology between the two sequences and because we have no precise knowledge of the sequence which was previously resident in this expression site and presumably displaced by the 117 BC as part of its activation. However, we can say that it must occur to the 3'-side of nucleotide -76 as this is the first discrepancy between the 117 BC and 117 ELC and probably to the 5'-side of the HinfI site (position +1) because sequences to the right of this site

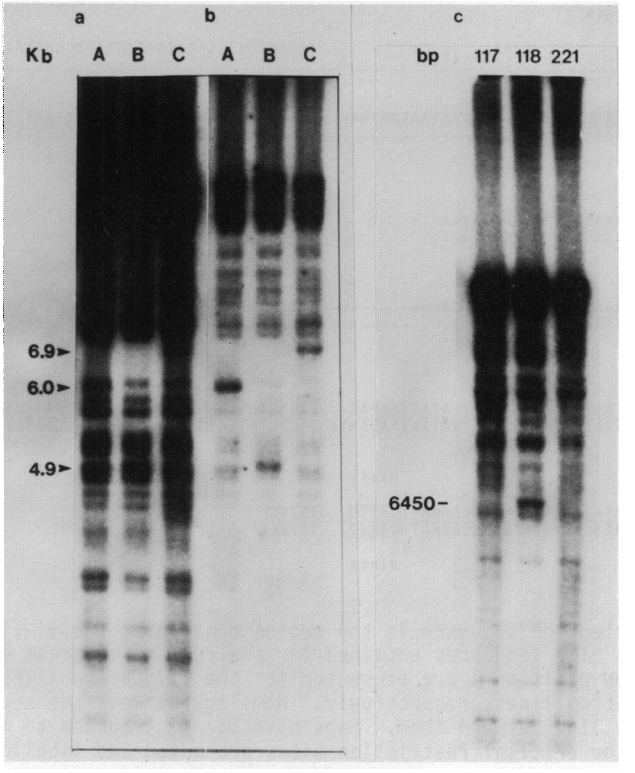

Figure 5. Detection of the 117 expression site in other variants using Southern blot analysis. About 1.5 µg of genomic DNA from three different variants, 117 (A), 118 (B) and 221 (C), was digested with HhaI, electrophoresed on a 0.7% agarose gel and transferred to a nitrocellulose filter. The filter was then probed with the HinfI-1207 fragment (see text) of pGE117.6 and washed at two stringencies: a. 0.1 X SSC at 37°C; b. 0.1 X SSC at 50°C. Sizes are in bp by comparison to bacteriophage λ DNA digested with PstI. c. A repeat of the experiment described in a, except the genomic DNA was digested with HindIII. Washing was with 0.1 X SSC at 50°C.

are specific to the 117 genes as judged by Southern blot analysis (see figure 2).

Characterization of the Barren Region

To investigate the structure and properties of the barren region, we undertook to first use fragments from pGE117.6 as ^{32}P-labelled probes in Southern blot analysis. The results of this experiment are shown in figure 5. In part "a" of this figure, hybridization with the HinfI-1207 probe (from positions +1 to -1207 of pGE117.6; figure 3) produces a smear on the autoradiogram indicating this sequence is highly represented throughout the

genome. At higher stringencies, however, discrete bands are detected in the three variants tested with, for the particular enzyme used, one band being unique to each variant (figure 5b). This unique band should represent the 5'-barren region of this expression site as it exists in other variants. To test this prediction, we repeated the hybridization using a HindIII digest of genomic DNA from three variants each expressing a different VSG gene. Published reports indicated that the 5'-barren region upstream of the 118 ELC (which uses the same expression site as that used in the 117 variant) is contained within a HindIII fragment of 6.3 kb (10). Figure 5c shows a unique band of about that size (6.45 Kb) in the 118 digest. The slight discrepancy in the sizes is within that expected for independent estimations by different groups. The corresponding fragment containing the 5'-end of the 117 ELC is about 10-15 kb (10; J.C.B., unpublished results) and could not be resolved from the less specific binding near the exclusion limit of the gel. No unique band could be detected in this experiment with 221 DNA and so we presume that here, too, the fragment containing the expression site used in the 117 and 118 variants is not resolved clear of the excluded material near the top of the gel.

Nucleotide Sequence of the Barren Region

Using the strategy shown in Figure 3, we determined the complete nucleotide sequence of pGE117.6 between positions +27 and -1589, representing the upstream barren region. Because of the paucity of suitiable restriction sites, it was necessary to subclone the insert in order to generate the complete sequence. This was done by cloning the 1.9 Kb AluI fragment (position +26 to a site within the vector, fig. 3) into the HincII site of the plasmid vector pUC8 (22). The resulting plasmid pGE117.61 was further manipulated by linearising with EcoRI, digesting with the double-strand specific exonuclease, Bal31, and religation in the presence of EcoRI linkers. This resulted in two plasmids, pGE117.611 and pGE117.612, as shown in fig. 3. All sites used for end-labelling and sequence determination were checked by examining sequence ladders generated from adjacent sites to ensure the contiguity of the sequence. Within regions sequenced on more than one plasmid (totalling about 350 bp, see fig. 3), no discrepancies were found testifying to the stability and fidelity of replication of these plasmids in *E. coli* despite the unusual and repetitive nature of the sequences present in the insert.

The sequence obtained for this region is presented in Figure 6. The portion containing the 5'-limit of transposition has already been presented

```
                                                                                                5'...GAT
-1586
CAGTAGTAATAATAATAATAATGATAATAATAATAGTAG---------GAGAGTGTTGTGAGTGTGTGTATATACGAATATTATAATGAGAG
-1503
CAGTAATGATAATAATAATAATGATGATAATAATAG------------AAGAGTGTTGTGAGTGTA-GTATATACAAATATTATAATGAGAG
-1424
CAGTAATAATAACAATAATAATAATAATATTAG---------------GAGAGTGTTGTGAGTGTG--TATATACTGATATTATAATAAGAG
-1349
AAGTAATAATAATAATGATGATAATAATAATAATAG------------GAGAGTGTTGTGAGTGTGTGTATATACGAATATTATAATAAGAG
-1269
CAGTAATAATAATAATATTAATAATAATAATAATGATGATAATAATAGAAGAGTGTTGTGAGAGTCTGTATATACGAATATTATAATGAGAA
-1177
CAGTAATAATAATTACAATAATAATGATAATAATAATAATAA------GAGAGTGTTGTCAGTGTGTGTATATACGAATATTATAATAAAAG
-1091
CAGTAATAATAATAATAATAGTAAGAA
                           TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
TAATAATAATAATAATAATAATAATAATAATAATAATAATAATAA
                                                GAGGAGAGTGTTGTGAGTGTGTGAATATACTAATATTATGATAAGAG
-747
CAGTAATAATAATAATAATAG---------------------------GAGAGTGTTGTAAGTGTG--TAAATACGAATATTATAATAAGAG
-684
CAGCACTAACAATAATAATAATAG------------------------GAGAGTGTTGTAAGTGTG--TAAATACGAATATTATAATAAGAG
-618
CAGCACTAACAATA-TAATAATAG------------------------GAGAGTGTTGTGAGTGTG--TATATACGAATATTATAATAAGAG
-553
TAGTAATTATAATAATAATAGTAATAATAG-------------------AAGAGTGTTGTGAGTGTGCATATATACTAATATTATAATAAGAG
-479
CAGTAATAATAATAATAATGATAATAATGATAG----------------GAGAGTGTTGTGAGTGTG--TATATACGAATATTATAATAAGAG
-404
CAGTAATAATAATAATAATAATAATAATAATAATAG---------GAGTGTGTTATGAGTGTG--TATATACGAATATTATAATAAGAG
-323
CAGTAATAATAATAGTAGTGATAATAATAG-------------------GAGAGTGTTGTGTGTGTG--TATATACGAATATTATAATAAGAG
-251
CAGTAATAATAATAATAATGATAG------------------------GAGAGTGTAGTGAGTGTG----TATACGAATATTATAATAATAG
-187
TAATAATAATGATAATGATAATAATAG----------------------AAGAGTGTTGTGAGTGTG--TATATACGAATATTATAATAAGAG
-118
CAGTAATAATAGTAATGATGATGTTAATAATAATAATAATAG------AAGAGTGTTGTGAGTGTGTG--TATACGAATATTATAATAAGAA
-34                 +1
TGATAGTTGTTTATACAGAATACAAACCAAAATTGAATCATTTAATACAATGGAAAGTGAG...117ELC
```

Figure 6. Nucleotide sequence of the upstream barren region as cloned in pGE117.6. The sequence is numbered from the HinfI site near the 5'-end of the 117TS. Gaps have been introduced to facilitate comparison of the repeat segments comprising this portion of the barren region. The $(TAA)_{90}$ stretch is presented as an insertion within the middle of one such repeat. Underlined sequences indicate the restriction sites shown in the map in fig. 3: -1589, Sau3A; -1208 and +1, HinfI; -685 and -619, Fnu4H.

in Figure 4. Like this sequence, the remainder of the barren region as cloned in pGE117.6 consists primarily of 76-bp repeats. This length is, in fact, a mean as these repeats vary in length between 63 and 92 bp. Their composition and basic motif, however, is highly conserved with virtually all differences conforming to the general motif (e.g., variation is usually in the number of TAA repeats found or substitution of G for A or C for T). It is interesting to note that the region from -736 to -605 contains two repeat units, the compositions of which are perfectly conserved with respect to each other, both in length and the presence of exceptions to the usual pattern. For example, the second triplet of these repeats is CAC (vs. the usual TAA) and immediately following the G/T rich region, they have TAAAT instead of TATAT. Both these exceptions are unique to these repeats and

probably represent a recent dupliation of one segment, as discussed below.

In addition to the 76-bp repeats, there is a striking sequence of 270 bp consisting of the TAA-triplet repeated 90 times with absolute fidelity. This sequence is found as though interrupting one of the 76-bp repeats, except that the usual TAAGAG is expanded to TAAGAA$(TAA)_{90}$GAGGAG.

DISCUSSION

5'-limit of Transposition

We have found that the 5'-limit of transposition of the 117 BC lies within or very near to the first of three 76-bp repeats. These repeats have also been found upstream of the 118 BC and the suggestion was previously made that they would contain the transposition limit (29). Given the findings of Van der Ploeg, et al. (30) which suggested that many, if not all, VSG genes have a similar sequence at the 5'-end of their transposed segment, it seems likely that the 76-bp sequence is required for the gene conversion event which gives rise to ELC's. The precise mechanism by which this occurs is not yet known and is not discernible from the sequence itself. One possible clue may be the presence of the sequence TGTTG which has been found near the ends of other eukaryotic transposable elements (31) as has been previously suggested by others (6).

The 76-bp repeats are exclusive to the 5'-end of the transposed segment, the 3'-end possessing a different and less-ordered homology block as already described (15). This asymmetric arrangement seems likely to ensure the correct polarity of insertion when generating an ELC.

The absence of the variant from which 117 was derived by a gene conversion event in this expression site prevents the otherwise desirable analysis of this expression site before and after insertion of the 117 VSG gene. We can, however, generate variants derived from gene conversion of the 117 ELC and these experiments, now in progress, should further delineate the sequences and mechanisms involved in the gene conversion.

Structure of the Barren Region

Figure 7 presents schematically the structure of that portion of the barren region which we have cloned in pGE117.6. We found no discrepancies between the maps obtained for this region as cloned and as it exists in the trypanosome genome as demonstrated by southern blotting. This suggests that the "barrenness" of this region is not due to modification of the bases in the trypanosome genome (32). Instead, it appears to be due to unusual, repeated sequences which because of their A/T richness and non-random

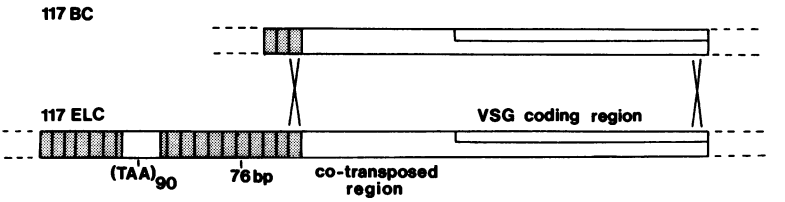

Figure 7. Schematic drawing depicting the organisation of the 117TS as it exists within the 117BC- and 117ELC-loci. The 117TS is presented within the two loci where it resides in the 117 variant. The 76-bp repeats are indicated by the shaded segments separated by the vertical lines. $(TAA)_{90}$ is shown within the upstream region of the 117 expression site. The 117VSG-coding region is boxed. The large crosses indicate the approximate limits of the transposition.

organization, lack restriction enzyme recognition sequences.

Given the repeated nature of this insert, we were concerned that its replication in E. coli might introduce changes due to recombination events despite the use of E. coli HB101, which contains the recA13 mutation. To test this, we compared the structure of pGE117.6 with a plasmid derived in a duplicate but independent experiment, pGE117.9. Detailed restriction analysis of the two plasmids gave identical results indicating no major rearrangements had occurred within the E. coli host.

The finding that the 5'-limit of transposition occurs within a sequence which is tandemly repeated many times in the 5'-barren region may explain the fluctuations in size which have been observed for this region (10). Such fluctuations could arise by variation in which 76-bp repeat unit (in both the BC and expression site) is targeted in the gene conversion. They could also be due to expansion and contraction of this region by intracistronic recombination events. Such an event seems likely to be the recent cause of the exact duplication seen by comparing the region between -730 to -671 with -670 to -605, as detailed above.

As with any tandemly repeated sequence, designation of the beginning and end of the basic repeat unit will vary depending on the criteria used. For the 76-bp repeats, we have chosen to consider the 3'-end of the repeat as the position where the consensus sequence is first encountered when working upstream of the transposed segment. Hence, the GAG (or GAA) triplet found at the right end of each line in figures 4 and 6 represents the 3'-end and the CAG triplet found at the left ends represents the 5'-end of each repeat (note that the sequence downstream of position -34 does not conform to the 76-bp consensus sequence). These designations are close to the end

points assigned by Liu et al. for the repeats upstream of the 118 BC (29). An alternative interpretation, however, is that the repeats begin at the GAG triplet found immediately following the $(TAA)_n$ region of variable length (i.e. immediately following the long dashed region in fig. 6). This interpretaion would place the $(TAA)_{90}$ insertion between repeats (rather than within) and the first difference between the 117 BC and 117 ELC (position -76) at the start of a repeat. Until the origin of these sequences is known we will not be able to discriminate between the two alternatives.

The reason for the unusual structure of these 76-bp repeats upstream of the ELC is not clear, especially given that transcription of these genes is seemingly not initiated within 30 kb upstream of this region (12,13). It has been previously pointed out (29) that these repeats share some properties with "enhancer" elements found in other systems. Yet, such sequences are not known to exert their effect in these other systems at the distances (i.e. $\geq$ 30 kb) involved here. What then is their function and how is it effected?

To understand this, we must first return to the problem of how the ELC is selectively transcribed over the BC and how one expression site is targeted for expression over the others. One model supposes that a promoter region lies upstream of an expression site and that very long primary transcripts are produced and then processed (33). The model presumes no promoter exists upstream of the BC genes and that different expression sites are made contiguous with the promoter region by chromosome-end exchange (remembering that all expression sites are located near such ends). Although this model must not be dismissed too readily, it is unsatisfying because of: 1. the wastage in such a transcriptional system if the primary transcript consists largely of a single, very long intron which is subsequently spliced out; 2. the inability to detect VSG-containing transcripts longer than about 5 Kb (12,34); and, 3. it does not explain the detection of mini-exon derived sequences in the RNA of cultured insect-forms of trypanosomes where no VSG gene expression occurs (12,13).

An alternative model which accounts for these observations is one in which RNA polymerases bind and initiate transcription at each of the 1.35 Kb mini-exon repeats. Transcription continues and terminates at a discrete site within each repeat resulting in the synthesis of a short transcript (i.e. less than 1.35 Kb). The polymerase would then dissociate from the mini-exon repeats, possibly with the short transcript in tow. Reinitiation of transcription might then occur upstream of an ELC using the short

transcript as a primer with the resulting transcript being spliced to give the final mRNA with a minimum of wastage. Regulation of expression could then be controlled at the level of reinitiation of transcription, with the same primer being utilized in the switching on of different expression sites. According to this appealing but speculative model, something must identify the activated expression site with its resident ELC as distinct from both the silent BC and other, inactive expression sites. Such a label could be a mobile reinitiation site which would move between expression sites. If only one such element exists, this would explain the mutual exclusion operating between expression sites.

One possible candidate for such a mobile reinitiation element is the $(TAA)_{90}$ sequence. It possesses three critical properties: first, it is flanked by 76-bp repeats known to be involved in transposition of a BC gene into an expression site. These sequences might also enable movement of $(TAA)_{90}$. Second, its insertion into an expression site might not be detected because, like the remainder of the 5'-barren region, it lacks restriction sites. Also, if $(TAA)_{90}$ is inserted by homologous recombination between the 76-bp repeats, no significant change in the size of the 5'-barren region need occur (the recombination might occur between 76-bp repeats number X and X + 4, i.e. about 300 bp apart). This might explain how some expression sites are apparently activated with no detected alteration in their structure (1, 35). Third, $(TAA)_{90}$ could provide an excellent reentry site for RNA polymerase because of its pure A/T composition (easy melting) and repeated structure (possibly leading to D-DNA formation; 36).

If the model is correct, it might explain the function of the 76-bp repeats as these could act as enhancers of reinitiation, rather than the usual role of enhancers in initiation at promoters (37). This is consistent with the proposed mode of action of enhancers whereby they facilitate the entry of RNA polymerase molecules which then initiate nearby (37). We are currently testing this model by investigating the primary transcripts of an ELC and further characterizing $(TAA)_{90}$ with respect to its abundance in the genome and mobility.

ACKNOWLEDGMENTS

We wish to gratefully acknowledge the technical and secretarial assistance of Ms. Deborah Thornton and Ms. Mary Ann Siri, respectively. We are also grateful to Lex Van der Ploeg for providing the original 117 BC

clone and Paul Orndorff for helpful technical suggestions. Finally, we welcome and acknowledge the exchange of information prior to publication with Drs. N. Agabian, P. Borst, J. Donelson and their colleagues. This work was supported in part by a grant from the American Cancer Society: IN 32 W.

REFERENCES

1. Borst, P. and Cross, G.A.M. (1982) Cell 29:291-303.
2. Englund, P.T., Hajduk, S.L. and Marini, J.L. (1982) Ann. Rev. Biochem. 51:695-726.
3. Cross, G.A.M., Holder, A.A., Allen, G. and Boothroyd, J.C. (1982) Am. J. Trop. Med. Hyg. 29:1027-1032.
4. Boothroyd, J.C. and Cross, G.A.M. (1982) Gene 20: 281-289.
5. Liu, A.Y.C., Van der Ploeg, L.H.T., Rijsewijk, F.A.M. and Borst, P. (1983) J. Mol Biol. 167:57-75.
6. 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.
7. Longacre, S., Hübner, V., Raibaud, A., Eisen, H., Baltz, T., Giroud, C. and Baltz, D. (1983) Molec. Cell. Biochem. 3:399-409.
8. De Lange, T. and Borst, P. (1982) Nature 299:451-453.
9. Williams, R.O., Young, R.J. and Majiwa P.O. (1982) Nature 299:417-421.
10. Van der Ploeg, L.H.T., Bernards, A., Rijsewijk, F.A.M. and Borst, P.(1982) Nuc. Acids Res. 10:593-609.
11. Laurent, M., Pays, E., Magnus, E., Van Meirvenne, N., Matthyssens, G., Williams, R.O. and Steinert, M. (1983) Nature 302:263-266.
12. De Lange, T., Liu, A.Y.C., Van der Ploeg, L.H.T., Borst, P., Tromp, M.C. and Van Boom, J.H. (1983) Cell: 34 891-900.
13. Nelson, R.G., Parsons, M., Barr, P.J., Stuart, K., Selkirk, M. and Agabian, N. (1983) Cell 34:901-909.
14. Cross, G.A.M. (1975) Parasitol. 71:393-417.
15. Bernards, A., Van der Ploeg, L.H.T., Frasch, A.C.C., Borst, P., Boothroyd, J.C., Coleman, S. and Cross, G.A.M. (1981) Cell 27:497-505.
16. Simpson, A.J.G., Sher, A. and McCutchan, T.F. (1982) Mol. Biochem. Parasitol. 6:125-137.
17. McDonell, M.W., Simon, M.N. and Studier, F.W. (1977) J. Mol. Biol. 110:119-146.
18. Southern, E.M. (1975) J. Mol. Biol. 98:503-517.
19. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol 113:237-251.
20. Boyer, H.W. and Roulland-Dussoix, D. (1969) J. Mol. Biol. 41:459-472.
21. Twigg, A.J. and Sherratt, D. (1980) Nature 283:216-218.
22. Viera, J. and Messing, J. (1982) Gene 19: 259-268.
23. Grunstein, M. and Hogness, D.S. (1975) Proc. Nat. Acad. Sci. USA 72:3961-3965.
24. Birnboim, H.C. and Doly, J. (1979) Nuc. Acid Res. 7:1513-1523.
25. Boothroyd, J.C., Paynter, C.A., Coleman, S.L. and Cross, G.A.M. (1982) J. Mol. Biol. 157:547-556.
26. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology, Grossman, L. and Moldave, K., Eds. Vol. 65. pp. 499-560, Academic Press, New York.
27. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87:107.
28. Boothroyd, J.C., Cross, G.A.M., Hoeijmakers, J.H.J. and Borst, P. (1980) Nature 288:624-626.
29. Liu, A.Y.C., Van der Ploeg, L.H.T., Rijsewijk, F.A.M. and Borst, P. (1983) J. Mol. Biol. 167:57-75.

30. Van der Ploeg, L.H.T., Valerio, D., De Lange, T., Bernards, A., Borst, P. and Grosveld, F.G. (1982) Nucl. Acids Res. 20:5905-5923.
31. Varmus, H.E. (1983) in Mobile Genetic Elements, Ed. Shapiro, J.A. pp.411-503, Academic Press, New York.
32. Raibaud, A., Gaillard, C., Longacre, S., Hibner, U., Buck, G., Bernardi, G., and Eisen. H. (1983) Proc. Nat. Acad. Sci. USA 80:4306-4310.
33. 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, Eds. Chater, K.F., Cullis, C.A., Hopwood, D.A., Johnston, A.W.B. and Woolhouse, H.W., pp. 207-233, Croom Helm, London and Canberra.
34. Van der Ploeg, L.H.T., Liu, A.Y.C., Michels, P.A.M., De Lange, T., Borst, P., Majumder, H.K., Weber, H., Veeneman, G.H. and Van Boom, J. (1982) Nuc. Acid Res. 10: 3591-3604.
35. Young, J.R., Shah, J.S., Matthyssens, G. and Williams, R.O. (1983) Cell 32:1149-1159.
36. Selsing, E. Arnott, S. and Ratliff, R.L. (1975) J. Mol. Biol. 98: 243-248.
37. Fromm, M. and Berg, P. (1983) Mol. and Cell. Biol. 3: 991-999.