J Mol Evol (1996) 42:500–511

JOURNAL OF MOLECULAR EVOLUTION

Sequence Divergence in a Family of Variant Surface Glycoprotein Genes from Trypanosomes: Coding Region Hypervariability and Downstream Recombinogenic Repeats

Mark C. Field,* John C. Boothroyd

Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA

Received: 4 October 1995 / Accepted: 25 December 1995

Abstract. The surface of the parasitic protozoan *Trypanosoma brucei* spp. is covered with a dense coat consisting of a single type of glycoprotein molecule, the variant surface glycoprotein (VSG). There may be as many as 1,000 genes for VSG within the genome of *T. brucei,* and the switch of expression from one to another is the phenomenon of antigenic variation. As an approach to understanding the evolution of VSG genes we have determined the genomic DNA sequences of the eight genes encoding the variant surface glycoprotein 117 (VSG) family. From these data we have observed a number of features concerning the relationships between these genes: (1) there is a region of high variability confined to the N-terminus of the coding sequence, and comparison of the sequences with the available X-ray diffraction crystal structures suggests that two of the most variable stretches within the N-terminal domain are present on surface-exposed loops, indicating a role for epitope selection in evolution of these genes; (2) the 29 nucleotides surrounding the splice acceptor site are absolutely conserved in all eight 117 VSG genes; (3) numerous insertion/deletion mutations are located within or immediately downstream of the C-terminal protein-coding sequences: (4) within 500 bp downstream of the insertion/deletion mutations are one or two copies of a repeat motif highly homologous to the recombinogenic 76-bp repeat sequences present upstream of many VSG basic copy genes and the expression-linked copy.

Key words: Evolution — Multigene family — Recombination — Trypanosome — Variant surface glycoprotein

Introduction

Trypanosoma brucei spp., the African trypanosome, is the causative agent of African sleeping sickness in humans; it affects several millions of individuals. The parasite life cycle involves obligate transmission through a mammalian host and an insect vector, the tsetse fly (*Glossina* spp.). While in the mammalian host the parasite dwells in the intertissue spaces and the bloodstream. The parasite is covered by a dense proteinaceous layer consisting of ~10^7 copies of a single protein, the variant surface glycoprotein (VSG). During infection the parasite population cycles, with each successive parasitemic peak characterized by the emergence of a serologically distinct dominant cell surface antigen (reviewed in Cross

**Present address:* Laboratory of Cell Biology, Department of Biochemistry, Imperial College of Science, Technology and Medicine, London, SW7 2AY, UK

Abbreviations: BC, basic copy; ELC, expression-linked copy; ES, expression site; GPI, glycosylphosphatidylinositol; indel, mutation where insertion or deletion cannot be discriminated; (k)bp, (kilo)base pairs; ORF, open reading frame; URS, upstream repeat sequence (also known as 76-bp repeats); UTR, untranslated region; VSG, variant surface glycoprotein. The nucleotide sequences presented in this paper have been submitted to GenBank with the following accession numbers; pSUB85, L31608; pSUB70C, L31607; pSUB70A, L31606; pSUB60, L31605; pSUB55, L31604; pSUB52, L31603; pSUB50, L31602 and the 117 basic copy (pGB117), L34415

Correspondence to: J.C. Boothroyd

1990a; Van der Ploeg 1990; Pays et al. 1994) determined by which VSG is expressed. Parasites bearing the new dominant VSG are eventually eliminated by the host's immune system, and this phenomenon of antigenic variation is absolutely essential to the survival of the parasite. In the field the rate of switching from one VSG to another is estimated as ~10^{-2}, but in laboratory strains the switch rate is significantly less than this, ~10^{-6} (Lamont et al. 1986). As has been noted by Vickerman (1989), this latter rate is so low that it falls within the rate of random recombination events, and therefore a large number of the reported DNA rearrangements associated with antigenic switching may have been created by a mechanism unrelated to that utilized by a natural population in maintenance of an infection.

Those VSGs that have been characterized are glycoproteins of approximately 55–65 kDa and are anchored to the parasite cell membrane by a glycosylphosphatidylinositol (GPI) membrane anchor (Cross 1990b). It has been estimated that there are on the order of 10^3 different VSG genes within the parasite genome (Van Der Ploeg et al. 1982), which are loosely clustered. These clusters do not reflect a simple evolutionary relationship, however, as in the case of the 117 VSG gene family the most closely related genes are dispersed throughout the genome (Beals and Boothroyd 1992a).

Expression of VSG genes occurs from a specialized telomere-proximal expression site (ES). The VSG gene present in an ES can be overwritten by a gene conversion process (Pays et al. 1983), resulting in the replacement of the ES sequence with the donor or basic copy VSG sequence (reviewed in Cross 1990a; Van der Ploeg 1990; Pays et al. 1994). Alternatively, a new ES may be activated or sequence exchanged between two ESs to allow the expression of a new VSG. The details of these mechanisms remain poorly understood. The ES itself is a large polycistronic transcription unit, and is always located close to a telomere. Approximately 1–2 kbp upstream of the VSG gene in the ES are a variable number of ~76-bp repeats, which have been mapped as containing the 5′-limit of conversion (Campbell et al. 1984). These sequences are also present upstream of the VSG basic copy genes (Liu et al. 1983), while the downstream transposition limit is close or within the 3′ end of the open reading frame (ORF) (Bernards et al. 1981; Liu et al. 1983). Transposition of only part of a VSG sequence into an ES, using sequence similarity within the VSG ORF to initiate recombination, has been demonstrated, leading to the production of a chimeric VSG (Longacre and Eisen 1986; Thon et al. 1989, 1990; Kamper and Barbet 1992; reviewed by Barbet and Kamper 1993). There is also evidence for the introduction of apparently nontemplated point mutations specifically within the ORF of an expressed VSG gene (Lu et al. 1993). However, as information within the ES is believed to be eventually lost, these processes do not provide a mechanism for the diversification of the basic copy genes.

It is the mechanism of diversification that we address in this paper. Clones containing the genes for a discrete family of VSGs closely related to VSG 117 have been isolated in this laboratory (Beals and Boothroyd 1992a) by use of a probe to the region upstream of the ORF. We have now determined the complete nucleotide sequences of eight members of this family, covering a total of ~20 kbp of the *T. brucei* genome. Our analysis has been directed to addressing a number of specific questions. First, we have mapped the regions of greatest variability within the VSG coding sequences. Associated with this we have been able to reconstruct a probable course of evolution for the group of genes. Second, we have analyzed the sequences surrounding the ORFs of this panel of VSG genes to gain insight into the mechanisms and pressures leading to diversification. Finally, we have reexamined and extended the data on the location of the retroposon mobile elements RIME and Ingi, previously shown to be present near to the 117 VSG genes (Beals and Boothroyd 1992a).

Methods and Materials

Isolation and Preparation of DNA. Subclones containing genes for members of the 117 VSG family have been previously described in detail (Beals and Boothroyd 1992a,b). Plasmids and cosmids were grown in *E. coli* DH5α and DNA was prepared using a Qiagen plasmid isolation kit (Qiagen Inc.) exactly as described by the manufacturer. DNAs were further purified by precipitation with 13% polyethylene glycol 4000 or by ultrafiltration using a Centricon 100 where necessary.

Generation and Analysis of Sequence Data. The 117 VSG genes contained on the pSUB plasmid series (Beals and Boothroyd 1992b) were sequenced by gene walking without further subcloning or manipulation. Additionally we also determined the nucleotide sequence of the 117 basic copy from pGB117 (Boothroyd et al. 1982, Fig. 1). Sequence data were obtained using dye terminator chemistry with Taq-cycle sequencing using custom primers and the products were resolved using a 373 DNA automated sequencer (PAN Facility, Stanford University) as described by the manufacturer (Perkin Elmer Inc., Foster City, CA). Raw data from the sequencer base calling output were converted to Word files (Microsoft Corp.) and initially assembled and aligned by eye. Completed DNA sequences were translated (all three frames) using the Navigator program (Perkin Elmer Inc.) and aligned using the Pileup program from the GCG package (Devereux et al. 1984). These initial alignments were used to proofread the sequence data by analysis of the electrophoretogram output from the 373 sequencer. Final sequences were realigned using Pileup or GeneWorks. Unassigned bases (i.e., a base was detected but could not be assigned with absolute confidence) are designated as ''N,'' and the corresponding unassigned amino acid as ''X.''

Phylogenetic gene trees were produced using PAUP (Phylogenetic Analysis Using Parsimony) V3.1.1 (Swafford 1993) using the Pileup alignments of the translated sequences in both branch and bound and exhaustive mode. Both algorithms produced the same topology with insignificant differences in the lengths of the internodes (see Figure 6 legend). The branch and bound tree was then bootstrapped (10^3 replicates) to assess robustness. Analysis of the similarity index of the translations of the ORFs was performed using the PlotSimilarity program from the GCG package, with a window of 20 residues (Devereux et al. 1984). The PlotSimilarity graphic output was scanned, copied into Adobe Photoshop (Adobe Systems Inc.), and manipulated for presentation. Sequences of the 117 (MITat 1.2), 221 (MITat 1.6), and 118 (MITat 1.5) VSG ELCs were retrieved from GenBank (accession num-

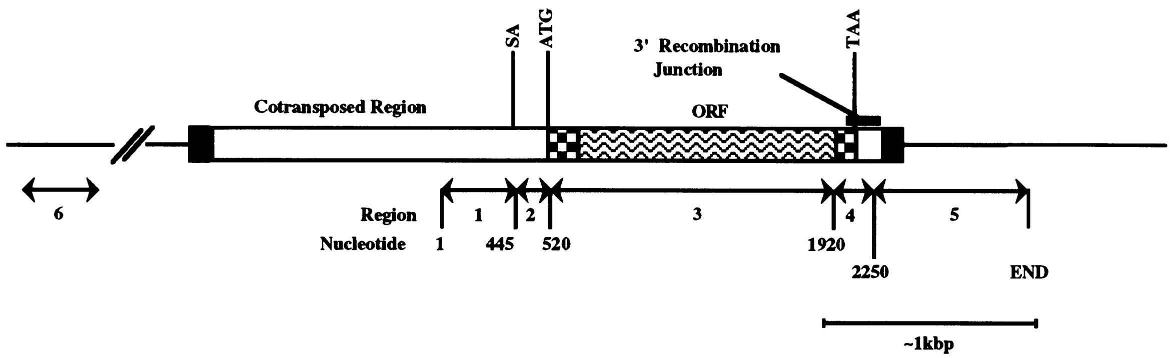


Fig. 1. Structure of a 117 VSG gene locus. The region considered as the locus is indicated as a *box,* and flanking sequence as a *single line.* The position of the URS is indicated as a *black box,* the N-terminal and C-terminal (GPI) signal sequences by *checker-boarding,* and the coding sequence by *wave-fill.* The proposed position of the recombination junction at the 3′ end of the locus is indicated by a *thick bar* above the schematic. Splice acceptor site, start codon, and stop codon are indicated by SA, ATG, and TAA (for the 117 basic copy), respectively. The positions of the regions 1–6 discussed in the text and the nucleotide positions of their boundaries are indicated beneath the schematic. Nucleotide numberings are based on pSUB70C and the alignments shown in Fig. 3 (not to scale).

bers K00638, V01387, K00639, X56762, and X56763). The sequence for pSUB60 was compiled using some previously determined sequence (GenBank accession number Z11676; Beals and Boothroyd 1992b) together with additional data generated in this study.

Southern Blot Analysis. Cosmid clones (pCOS85, 70A, 70C, 60, 55, 52, and 50), as described in Beals and Boothroyd (1992a), were digested with restriction endonucleases (New England Biolabs) and the fragments were separated in a 0.75% agarose gel. The fragments were transferred to 0.45-μm nylon membranes under alkaline conditions (modified from Maniatis et al. 1985) and probed with synthetic oligonucleotides using the fluorescein-terminal transferase system (Amersham International Inc.) and detected using the ECL reagent (Amersham International Inc.). All Southern blots were washed at high stringency (0.1 × SSC, 0.1% SDS, 65°C) for at least 2 h before visualization.

Oligonucleotides. Oligonucleotides were synthesized on an Applied Biosystems Oligonucleotide Synthesizer (Perkin Elmer Inc., Foster City, CA) and were used without purification. Sequences of the oligonucleotides used in Southern blotting were GGTCCAGTACCCC-GTATCATCGGGGGAAGCCAAGAGCCAGC (RIME1), GGCGCG-GCCATCAGCCATCACCGTA (RIME2), CGCCCCGCATGCT-CAACGCTCGAACAACTCCTGCACGTCCCG (Ingi1), and CACTGCGGTCTAGCGGCGACCCCG (VSG5.1). The RIME1 sequence was designed against part of the sequence from Hasan et al. (1984), while RIME2 was designed against a sequence determined for a RIME element on pSUB9.8 (T. Beals and J.C.B., unpublished data).

Results and Discussion

Location of RIME and Ingi Elements

It was previously suggested that the occurrence of the retroposon-like RIME and Ingi elements (Kimmel et al. 1987; Smiley et al. 1990) on some of the cosmids (from which the pSUB series of plasmids were derived) containing the 117 VSG family member genes (pCOS9.8A, 9.8B, 8.5A, and 6.0, Beals and Boothroyd 1992a) could provide recombination sites allowing both the duplication of VSG sequences and the movement of existing sequences within the *T. brucei* genome. We investigated this possibility by Southern analysis of the remaining cosmids using RIME and Ingi probes. Initially we were able to confirm the assignment of the VSG gene to the regions of the cosmids with homologous restriction maps with the oligonucleotide probe VSG5.1,, complementary to the 5′ end of the 117 VSG ORF (Fig. 2). We next probed with RIME1, RIME2, and Ingi oligonucleotide probes (see Materials and Methods). We were able to confirm the assignment of RIME/Ingi-related sequences on pCOS9.8A, 9.8B, 8.5A, and 6.0, as well as to detect additional elements in pCOS70C and 6.0 and two in pCOS5.0 (Fig. 2). Interestingly we were able to detect some of the RIME elements with the RIME1 oligonucleotide and others with the RIME2 probe, suggestive of some degree of heterogeneity within the RIME population.

The locations of the putative RIME/Ingi elements were not consistent with a role in duplication/transposition of the 117 VSG gene family; i.e., they were not localized at constant positions with respect to the VSG coding sequence (Fig. 2). The number of RIME/Ingi elements in the *T. brucei* genome is estimated at ~1,000 (Hasan et al. 1984), and therefore it is not unexpected that one or more RIME elements would be present in random DNA fragments of several tens of kilobases, i.e., the size of a cosmid insert. In addition, two of the cosmids analyzed, pCOS5.5A and 5.2A, did not contain RIME/Ingi-related sequence, despite containing sufficient DNA to cover regions corresponding to the positions of RIME/Ingi elements detected on the other cosmids. Taken together we conclude from these data that there is no evidence for a RIME/Ingi-mediated mechanism for expansion of the VSG gene repertoire.

Nucleotide Sequences of the 117 VSG Gene Family

The complete nucleotide sequences of eight 117 VSG gene family members were determined and are available from GenBank using the following accession numbers:

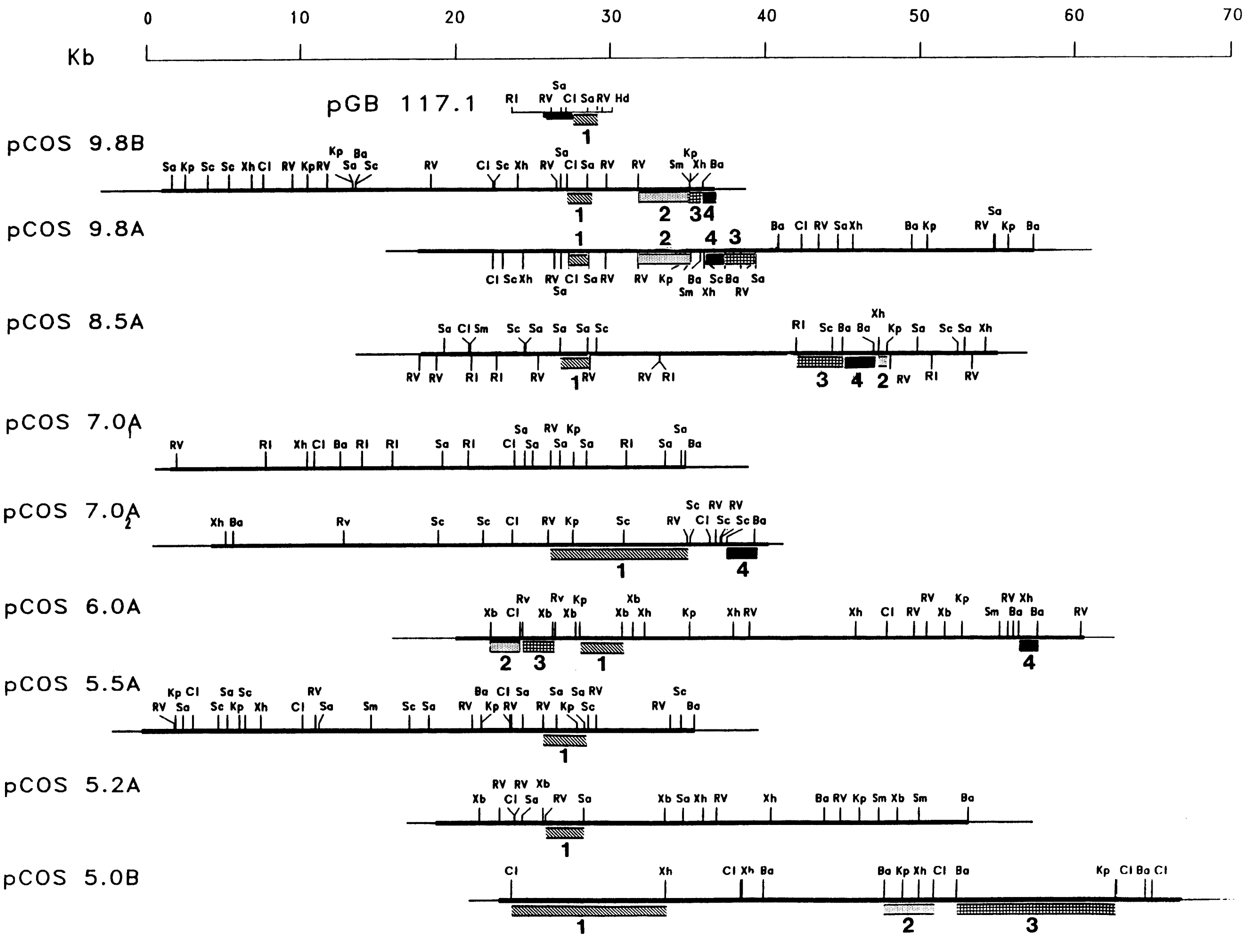


Fig. 2. Lack of correlation between the locations of RIME and Ingi elements and the 117 VSG gene family coding sequence. The restriction maps of the cosmid clones containing the 117 VSG family genes are shown. The positions of fragments which hybridize to VSG coding sequence (*box 1*), RIME1 (*box 2*), RIME2 (*box 3*), and the Ingi elements (*box 4*) are indicated (see Materials and methods). The restriction maps are redrawn from Beals and Boothroyd (1992a). pCOS 7.0_1A was not analyzed.

pSUB85, L31608; pSUB70C, L31607; pSUB70A, L31606; pSUB60, L31605; pSUB55, L31604; pSUB52; L31603; pSUB50, L31602; and the 117 basic copy (pGB117), L34415. The biologically relevant regions are shown in Fig. 3 following alignment.

A number of features are immediately obvious by inspection of the nucleotide sequences. First, the VSG gene sequence contained on pSUB70A and pSUB70C are extremely similar. The possibility that these two genes are allelic seems unlikely given the fact that all the other family members are present apparently once per diploid nucleus and the dissimilarity of the restriction maps of the respective parental cosmids (Beals and Boothroyd 1992a). It seems more likely, therefore, that they are the result of a very recent gene duplication event.

Second, there is, as expected, a high degree of similarity between all of the sequences throughout most of the analyzed region. We observed very few insertions or deletions in the sequences, i.e., the majority of the differences are due to base changes and not changes in the number of bases. This is most clear in the region from the start codon at nucleotide 520 through to the extreme C-terminal region following nucleotide ~1,900. This ensures that the reading frames of the VSG genes are maintained (except for pSUB70A and pSUB70C as discussed below).

The sequence of the basic copy presented here was determined from the genomic DNA contained on plasmid pGB117. Comparison of this sequence with the cDNA reported by Boothroyd et al. (1982) demonstrated essentially complete identity, with the exception of seven nucleotide differences. Of these only two are coding, changing amino acid 231 from tyr to phe and amino acid 268 from arg to gly. Point mutation differences between an ELC and a basic copy confined to the coding sequence have been reported recently (Lu et al. 1993), but note especially in this specific case that the MITat1.4 parasite preparation used for the mRNA (i.e., cDNA) cloning was different from that used for the genomic cloning: The two were separated by several animal passages over several years. Hence, it is not possible to comment on the speed of these changes.

The remaining discussion will focus on six regions (designed 1–6; see Fig. 1 for details). The regions will be considered in numerical order.

Region 1 is the upstream region not present in the mature VSG mRNA (nucleotides 1–444). As originally reported in Beals and Boothroyd (1992b), this region shows very high homology, among all the 117 VSG family. In this case we have not obtained sequence data as far upstream from the splice site as in that previous report, but we have seven examples over this region (but see below). Our alignment is slightly altered from that reported earlier (Beals and Boothroyd 1992b) due to the presence of additional sequences in this region, especially the closely related pair 70A and 70C. We find that region 1 is very highly conserved, and the single peak of variability at nucleotide ~250 can be accounted for by an eight-base region of unusual heterogeneity (nucleotides 253–260), including a single base insertion in the basic copy at nucleotide 258 and an eight-base difference between pSUB70A and pSUB70C and the remaining genes. We do not believe that this small region of heterogeneity has functional significance. It is worthwhile noting that the level of similarity is not particularly enhanced more than 29 nucleotides upstream of the splice acceptor site (compare nucleotides 101–200 with 321–420), even though this region may be more constrained than sequences further upstream due to requirements that it retain the ability to be a substrate for the *trans*-splicing machinery (e.g., branch formation). However, around the splice site itself we observe the absolute conservation of 29 nucleotides ACCTCCAACATAAGCAG/CAAAAGACTAGA (splice site for the 117 basic copy is indicated by ‘‘/’’; Boothroyd and Cross 1982) in all eight of the genes. This suggests that these 29 bases are required for correct recognition and processing of the nascent VSG message. It is important to bear in mind, however, that the sequence surrounding the splice sites of non-VSG genes in *T. brucei* is divergent from this sequence, and therefore this complete conservation may indicate a function for this region beyond being a substrate for the *trans*-splicing machinery.

Region 2 encodes the 5′ untranslated region (UTR), i.e., from splice acceptor site (nucleotide 445) to the start codon (nucleotide 520). We confirm the observation of Beals and Boothroyd (1992b) that over this region the basic copy is less related to the family members than they are to each other, consistent with the basic copy as an outlying gene (see below) and with there being less functional constraints on this region than that immediately surrounding the splice site itself.

Region 3 spans the open reading frame (nucleotides 520 to ~1920) to the GPI-addition site near the C-terminus. We observe several areas of increased variability within the N-terminal region of the open reading frame; these are discussed in more detail with respect to the protein coding function below. Within the C-terminal region, before reaching the sequences specifying the GPI-signal peptide, the levels of similarity are high. There are a small number of apparent deletions in some of the sequences as revealed by the alignment, specifically between nucleotides 688 and 711. However, there is no loss of reading frame through this particular region as all of the genes have a single base missing within this 23-base region relative to the consensus.

Region 4 (nucleotides 1920–2250) includes the GPI-signal peptide and approximately 200 bases of 3′ UTR (it is not possible to predict polyadenylation sites in *T. brucei* but this is the approximate size of the 3′ UTR in most VSG cDNAs so far identified). The nucleotide alignment reveals a different situation in region 4 compared to the others (Fig. 3B). Beginning downstream of nucleotide 1900 several large gaps need to be introduced to maintain the alignments, which is within the C-terminal region of the open reading frame. It is also necessary to introduce gaps within the region 3′ to the stop codons (nucleotide 2094 in the basic copy), and alignment without gaps only becomes possible again around nucleotide 2340 (see below).

It is also apparent that the alignment overall within this region is relatively poor, although it is still significant, and is in stark contrast to the high degree of homology that is seen immediately upstream within the C-terminal region of the ORF (compare the first block of nucleotides in Fig. 3B with the remainder). This increased heterogeneity does not mark the end of the conserved sequences as further downstream, in region 5, close similarity returns. This is consistent with the observation of significant similarities in the restriction maps of DNA downstream of and including region 4 for the cosmids containing the 117 VSG genes (Beals and Boothroyd 1992a), and leads to the conclusion that region 4 contains a localized area of highly increased sequence variation, characterized by both low homology at the nucleotide level as well as the presence of indels (mutations that are insertions or deletions, but due to a lack of temporal information it is not possible to conclude which). This type of heterogeneity is totally unlike that observed within the 5′ end of the ORF, which is characterized by the occurrence of point mutations but seldom an indel and contrasts dramatically with region 1, which has no obvious functional constraint on sequence divergence yet is extremely highly conserved.

This type of heterogeneity could be the result of inaccurate repair to a DNA duplex involved in a recombination event, e.g., transposition or gene conversion into an ES. Incomplete homology between the incoming VSG and that present in the ES would result in the loss or acquisition of sequence information within the donor due to inaccurate resolution of the recombination intermediates, manifested as indels within an alignment of otherwise-homologous genes. The indels could also be the result of an aborted recombination event where a donor gene failed to gene convert the ELC, and the resulting mutations are the product of a DNA repair mechanism. Because VSG genes are present within clusters (Van der

```
A
        101                                                                                                       200
70C     GCCAACTGTA GCGCGGCAAG AGGATCAAAG CGGCTTGGTA TTTTACATTT TTTTTTGCAG ACGACAGGCT GTGTGGCGGA CTAAAGCCGT CAGCAGACGA
70A     A...G..... .......... ........G. .......... .....G.... .......... G......... .........G .......... ...G......
85      A.....C... TG........ ......G... .....C.... .......--. .......... .AA....... ..A....... .....A.... ..........
52
50      A...-..... TG..-..... ......G... .....C.... .......--. .......... .AA....T.. ..A....... .....A.... ..........
60      -....AC... .G...-.G.. ....G.G... ....G..... .......--. .......G.. ..A.G..T.. ....T..AA. .....A.... ...G..G..T
117     A.....A.C. .TA....... ......G... .......... .......--. .......... ..A....A.. ..A....A.. .....A.... ..........
55      A.....C... .......... .....AG... .....C.... .......--. .......... .AA....... ..A....... .....A.... ..........

        201                                                                                                       300
70C     GTGTTCTTAT TCCGACCAGA AAACGAACCT GGCGAAGCGA AGGCATAACT TGTGTCC-AA AGCTCCAGCA AGGCTGTAAG CAATCCGAAC AAGTAAAACC
70A     .......... .......... ..G....... .......... .......... ...TC....G .......... .......... .......... ..........
85      .........A ..G..A.... .......... ......C... .......... ...CCT...T .C....TCTG G.AA.....C .......... .....C....
52
50      .........A ..G..A.... .......... ......C... .......... ...CCT..GT .C....TCTG G.AA.....C A......... .....C....
60      .....G...A ....-A.... ..G....T.. ...T.GC... ........T. ....CT...T .C....TCTG GCTA.....C ....G..... G..C.C...G
117     ..T......A .......... .......... .......... ........A. ....CT.A.T .C....TCTG G.AA.....C .......... .....C....
55      .........A .......... .......... .......... .......... ....CT...T .C....TCTG G..A.....C .......... ..........

        301                                                                                                       400
70C     AGGTCAATAA ACTCTACCTT TCGTAAAAGA GGCCACCGCG TCAGAAGCTA AGCAGCA--A CCTCC-GCAA ACAATGCCGG ATGACCCAAG GCAAGGCCAC
70A     .......... .......... .......... .......... .......... .......... .......... ....A..... .......... ..........
85      .......... .......A.. .......... .....G.... C......... .....G.GC. .......... ....A..... .....G.... ...C..A...
52                                                                                                ......A... ...C..A...
50      .......... .......A.. .......... .....G.... C......... .......GC. .....A.... ....A..... .....G.... ...C..A...
60      .......... .....G.A.. .......... .....G.... .G........ .......... .......... ....A..... .......... ......A...
117     .......... .......... .......... .....G.... C......... .......... .......... ....A..... .....G.... ...C..A...
55      .......... .......... .......... .....G.... .......... .......... .......... ....A...A. C....G.... ...C..A...

        401                                                                                                       500
70C     CGTACTCGAC ATTCACCCCG AACATTCACC TCCAACATAA GCAGCAAAAG ACTAGAA--A AAAACAAAAT ATTTACCAAA GAACAACTGG GCTTCAACAG
70A     .......... .......... .......... .......... .......... .......A.. .......... ..A....... .........C ..........
85      ....T.T... .......... .....C.... .......... .......... .......GC. -.C-...C.C ..A....... ..GA...G.. .T.......A
52      ......TA.. .......... .......... .......... .......... ......GGC. -.C-...C.C ..A....... ..GA...G.A .T.......A
50      ....T.T... .......... .....C.... .......... .......... .......GC. -.C-...C.C ..A....... ..GA...G.. .T.......A
60      ....T.T..G ..A....... .G...C.... .......... .......... ......GAC. C.C....C.T ..A..G..C. .-----...T ...G..T...
117     ...C..TA.. .......... ......T... .......... .......... .......GC. .--G..GCGC ..A..G.GC. A.CA.T.G.. .T.......A
55      ......T... ......T... .......... .......... .......... .......AC. .-C-...C.C ..A....... .T.A...GT. .T.......A

        501                                                                                                       600
70C     AAACGACTAC G-CTCGCGAA TGGATTGTCA GAACCGCGCA G-AAATAGCA CTGGTACAGT GGAAAGCAGC CACAGTAGC- GGCAGTGGCG TTGCTCTACG
70A     ......A... .......... .......C.. .......... .......... .......... .......... .......... .......... ..........
85      ......A... .......... .......C.. .......... ..C....... ......A... .......... .....C.... .T........ ..........
52      ......A... .......... .......C.. ........T. ..C....... ......A... ..N...GCAG .....C.... .C........ ..........
50      ......A... .......... .......C.. .......... ..C....... ......A... .......... .....C.... .......... ..........
60      ...A..A... ......G... .......C.. .G.....T.. ..C....... .......... .......... .....C.... .T........ ..........
117     .....GGAG. .A...A.-.. ....C..C.. -T..AAAGG. .AC.C...GG G.CAC...A. ..-.G..GAT ..AC.AT..T AA..C.AT.A C....T....
55      ......A... .......... ......TC.. .......TT. ..C....... .......... .......... .....C..-T .......... ......C...

        601                                                                                                       700
70C     TGGCCGAGAC AGCATCAGGA AGCTACGACG CCCTCGAATA CACAACTTGG TAAACTCACT GCGGTCTAGC GGCGCACCCG AGAAAGGTCA CCCGGAGGAG
70A     ......C... .......... ........A. .......... .......... .......... .......... .....C.... ........-. ..........
85      ......T... .......... ........A. .......... .......... .C...C.... .......... .....C..T. ........A- .A......TA
52      .A....T... .......... ........A. .......... .......... .C........ .......... .....C.... .........- .A........
50      ......T... .......... ........A. .......... .......... .C........ .......... .....C.... ........A- .A......CA
60      ......T... .......... ........A. .......... ...G...... .C........ .....G.... .....C..T. ..G.....A- ..T.....CA
117     CCAT.ACTC. ...GGAC..C GC.A.A..A. ....T..... ..A....... AC..AC.... ....A..G.. ...C.CA.T. ........TG .-...T....
55      ......TT.. .......... ........A. .......... .......... .C........ .......... .....C.... ........A- ..........

        701                                                                                                       800
70C     TGTTGGCAAA -CTGGAAAGT CAAACGAGCT ACAGCAACAA ACTGGAAGAA ATGGGGGAAA AGTTACGATT CTACGGCCTA AAAGGAGCAG CAGGAGGCGA
70A     .......... C......... .......... .......... .......... ....A..... .......... .......... .......... ..........
85      .A..A..G.. A......... ....TT.... ..C....... .......... ....A..C.. ..C....... ......A... ...A...... G...C.A...
52      .A..A..G.. A......... ...TTT.... .TC.A..... .......... ....A..C.. ..C.....A. ......A... ...A...... ....C.A...
50      .A..A..G.. A......... ....TT.... ..C....... .......... ....A..C.. ..C.....A. ......A... ...A...... G...C.A...
60      .A........ A......... ....T..... ..C....... .......... ....A..C.. ..C.....A. .......... .......... ....C.A...
117     .A..AA.G.. A...A....C ..C.TT.... ..C.G..A.. .......... ....AAACG. ..C.....A. .....CA... .......AC. G..TG..A..
55      .A..A..G.. A......... ....T..... ..C....... .......... ....A..C.. ..C.....A. ......A... ...A...... G...C.A...

        801                                                                                                       900
70C     GCAAACTACA GGCGATATGC TGGCATCAGC CGCAGCGCTG AAACAAATG- GCTGTTCAGG AAAAAGAACA AATCAATATG CAAACTGCGG CCGCGGCCGC
70A     .......... .C........ .......... .......... .......... .......... .......... .......... .......... ..........
85      ...G...... .T........ .A......A. .........C .TG.G...-G ...AGC.TCA ..G.....A. ..C....... .....C...C TTATA....T
52      A.GG..C... .TA..GC... ....CAA.A. G..G.....C .TG.G.G..A ..AAG-..AA .C.....G.. ..C....C.A ..G..G.... TTA.A....T
50      ...G...... .T........ .A......A. .........C .TG.G...-A ...AGC.TCA ..G.....A. ..C....... ....C.C..C TTATA....T
60      ...G...... .T........ .A......A. .........C .TG.G....G -..AGC.TCA ..G.....A. ..C....... .....C...C TTA.A....T
117     ...A.AAT.. .CG..G..A. .A...A..A. G.....C..A .TG.G.-CAA AAA.C..TCA C.CC....G. .GC....T.. A....A...C TGAA...G..
55      ...G...... .TA..GC... .......... .......... .......... .......... .......... .......... .......... TT........
```

Fig. 3. DNA sequence of the VSG 117 basic copy and the corresponding sequence from seven VSG 117 gene family members. *117,* VSG 117 basic copy; other numbers represent individual family member genes defined by the size of their *Hin*dIII fragments (Beals and Boothroyd 1992a,b). **A** Cotransposed region, 5′ UTR and N-terminal portion of the ORF. Initiation ATG is found at base number 520 and is *underlined.* The 29 nucleotides surrounding and including the splice acceptor site are *underlined* in pSUB70C. **B** 3′ portion of the ORF, 3′ UTR, and URS. Probable stop codons are *underlined.* The 15-bp conserved element is *underlined* for pSUB70C. The consensus sequence for the URS motif is shown below the corresponding regions of the gene sequences. Unassigned bases—i.e., a base is clearly present but the signal is ambiguous—are designated by *N.* Gaps introduced in the alignments are denoted by ‘‘−’’, and identity with the sequence of pSUB70C is shown as ‘‘.’’. Sequence of the URS is adapted from Campbell et al. (1984).

Ploeg 1990) it is important to bear in mind that there may be a further VSG gene located downstream of the 117 VSG genes that we have analyzed, and therefore some of the indels that we observe here could be the result of recombination events upstream of such a VSG gene. It is also possible that the mutations in region 4 are ‘‘imported’’ from the ELC; i.e., after generation of an ELC, including acquisition of a new 3′ end (which is known to be a frequent event; Barbet and Kamper 1993), gene conversion using the ELC as the donor fixes the altered 3′ end in the repertory of stable VSG basic copy genes. There is as yet no evidence for such a ‘‘backward’’ event.

The conserved 15-bp sequence TGATATATTTTAA-

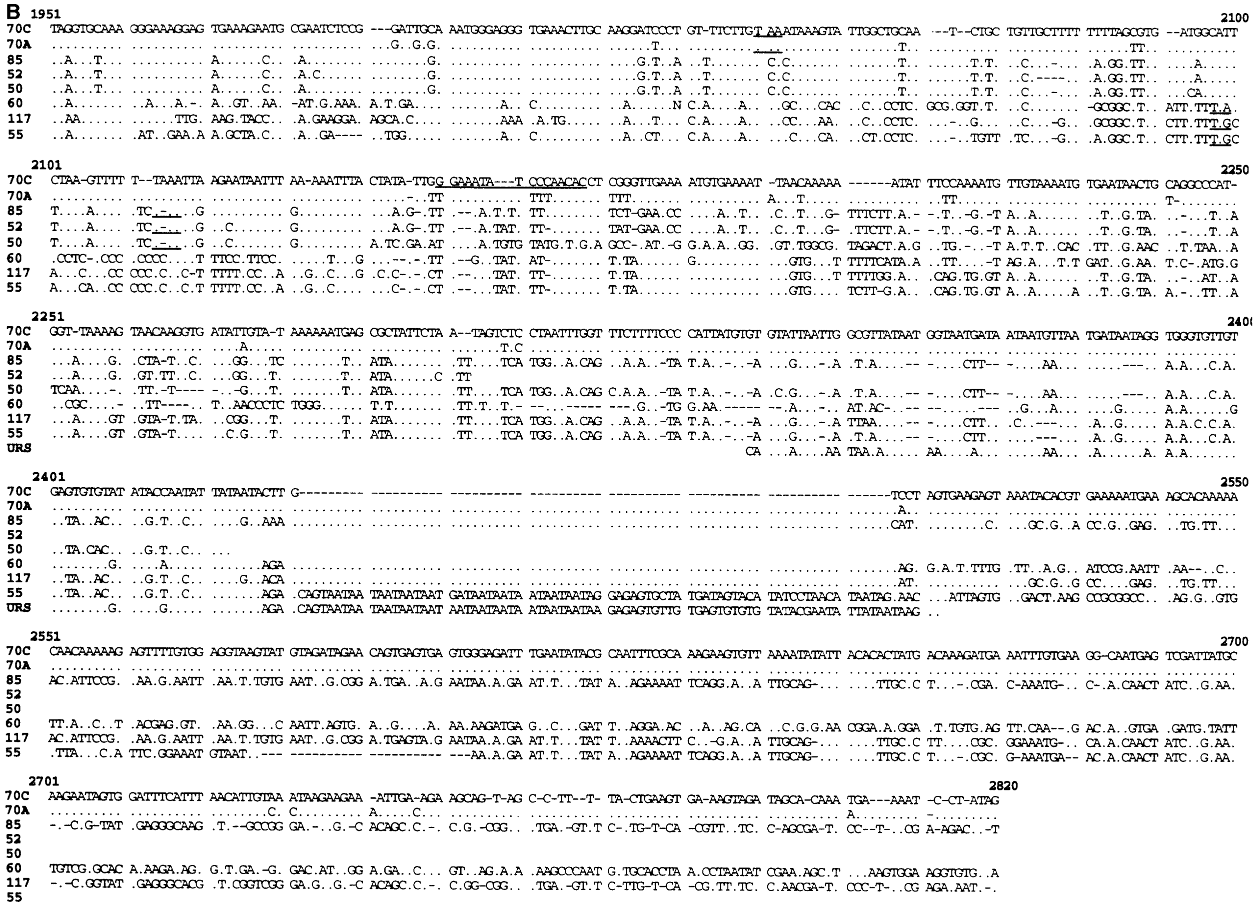


Fig. 3. Continued.

CAC, reported as present in the 3′ UTR of most expressed VSG genes (Borst and Cross 1982), occurs, with minor variations, in six of the 117 genes analyzed. Beginning at nucleotide 2150 this motif is conserved with one (basic copy, pSUB70A, pSUB55, and pSUB52), two (pSUB60), or three (pSUB85) changes. However, in pSUB70C and pSUB50 the motif is degenerate, with five and seven changes, respectively. As the role that the 15-bp sequence plays in VSG expression or recombination is not known, it is not possible to draw any conclusions from this observation. It is curious that the 15-bp motifs are highly divergent between pSUB70A and pSUB70C in an otherwise highly homologous stretch of nucleotides between these two genes.

Region 5, downstream of nucleotide 2250, contains sequence matching almost perfectly the consensus 76-bp upstream repeat sequences (URS) that have previously been seen in small numbers upstream of chromosome-internal VSG genes (Fig. 1) and in massive numbers upstream of VSG genes located in the telomeric ES (Campbell et al. 1984; Van der Ploeg 1990; Cross 1990a). One family member gene, pSUB55, contains two copies of the motif, whereas the other genes have a single copy. The second copy of the motif is present in pSUB55 as an insertion relative to the alignment with the other sequences, beginning at nucleotide 2431. This insertion sequence corresponds exactly to a single copy of the URS motif, beginning with a CAG nucleotide triplet. Most changes from the URS consensus are transitions. pSUB55 also has a 26-bp deletion downstream of the two URS motifs at nucleotide 2577. Homology between pSUB70A and pSUB70C with the rest of the genes decreases substantially after the insertion in pSUB55, but significant homology is retained between pSUB55, pSUB60, pSUB85, and the basic copy.

The function of these sequences close to the 3′ end of the VSG ORF is unknown, but the repeats upstream of the 117 BC ORF (at least) are involved in the recombination event into the ES (Campbell et al. 1984). It has also been suggested that the URS sequence can adopt a non-B-type helix—specifically, Z- and/or D-helix (Campbell et al. 1984). The junction of a left- and right-handed helix will result in the formation of a bulge, which would clearly have the potential to be recombinogenic.

Whether serving as the upstream or downstream recombination site the close proximity of the URS to the putative recombination junction may indicate a role as a recognition site for the recombination machinery, either acting as an entry point for an invading strand or providing a site that signals the downstream boundary of the recombinogenic region, i.e., preventing strand exchange

```
85 TTCTTGACGTACAATTTTTTCACTAAAGAAAAAGTAGCAAAGCTGATGACAGAATAAAAACAAACTTGCTTCCTGACTCTATTTTACCAAGCATTGCTCTTCACGCTGGACTCACT
60 ..................................C.............G...G.C..........C.TT..............C..T.............................
55 ............T......G..T...........C..AC.........G...G.C..........C..T..............C................................

85 TCTCTAAATAAGGCCCAAGAATCAGCGGTGGAAAAGACCAGCTAATTCGTGTTTAGGTTCGGCGTTAGCTACATTACAGGGGGCGACTGGCAGAACAAGCGGCTATTAAAAGACTG
60 C.........G....T............C.C.C....G.......G..........C...........A..........C........AA...G......T..C............
55 ..........G..G.T..AC........C..CC............G.....................T.........G.T..A......A...GG......AC..........T..

85 TGCTCATTTACTCTTATAAATTGATTCAAAGCAAATAGGGAAGCAGTTAACAGTAGCACAAAGACACCAAACAAACTGTTGATTAATGACATCAATATCAGCATCAACCTCTACAC
60 ...................C....C.......................................T.......................G....G......A....GG........G
55 ...................C....C.......................................T..................G.........G...................G.G

85 TAGCCAATAACCCTTTTCATTGTGAGCTCAGCT-GTCCATGCACAAAAAGGTGTCAAACACA-ACGCATGCGCCAACGAAT-GCCAGTGCAAGGCACGTCTTCAGACTCGGCTACA
60 CTAG..................C.......T............G......AG........ACGGG......AA..G.....T......................A.AG........
55 GTAA..G.................GC.........................G.C.........G.......AG..G........................................

85 GCACTATAGAGACATGGTTCCACGAGGCCAGCAACGC
60 ............G.....C.-............G...
55 .......G.......T....AC...............
```

Fig. 4. High degree of homology upstream of the VSG open reading frame in region 6. Sequence data were obtained using an oligonucleotide primer previously determined by sequencing the upstream region of the insert in pSUB85 from the vector. Sequence information could not be obtained from the remaining pSUB series plasmids because the inserts do not extend out to this region. Periods indicate identity to pSUB85 and ''–'' indicates a gap introduced in the alignment.

from taking place further downstream. Our demonstration that the 117 BC gene is in fact flanked by 76-bp motifs is highly suggestive that these sequences delineate a ''mobilization'' boundary, i.e., provide entry and exit sites for the recombination machinery on either side of the VSG coding region.

We also obtained sequence data from ~3 kbp upstream of region 1 for pSUB85, 60, and 55 (region 6). These sequences (Fig. 4) have a remarkably high degree of similarity even at this distance from the VSG ORF. Whether this similarity extends to all the family members could not be tested because the other pSUB series plasmids do not extend far enough upstream. These results are particularly significant because preliminary results indicate the presence of degenerate URS sequences ~1 kbp upstream of the VSG ORF in pSUB85, pSUB70C, and pSUB65 in a position analogous to where they are found in the 117BC gene (data not shown but see Fig. 1 for position). Although these URS elements were technically difficult to sequence, it was clear that they were much more divergent from each other than the corresponding sequences in region 6, again suggesting that the URS may be involved in local sequence variation similar to that seen in regions 4 and 5.

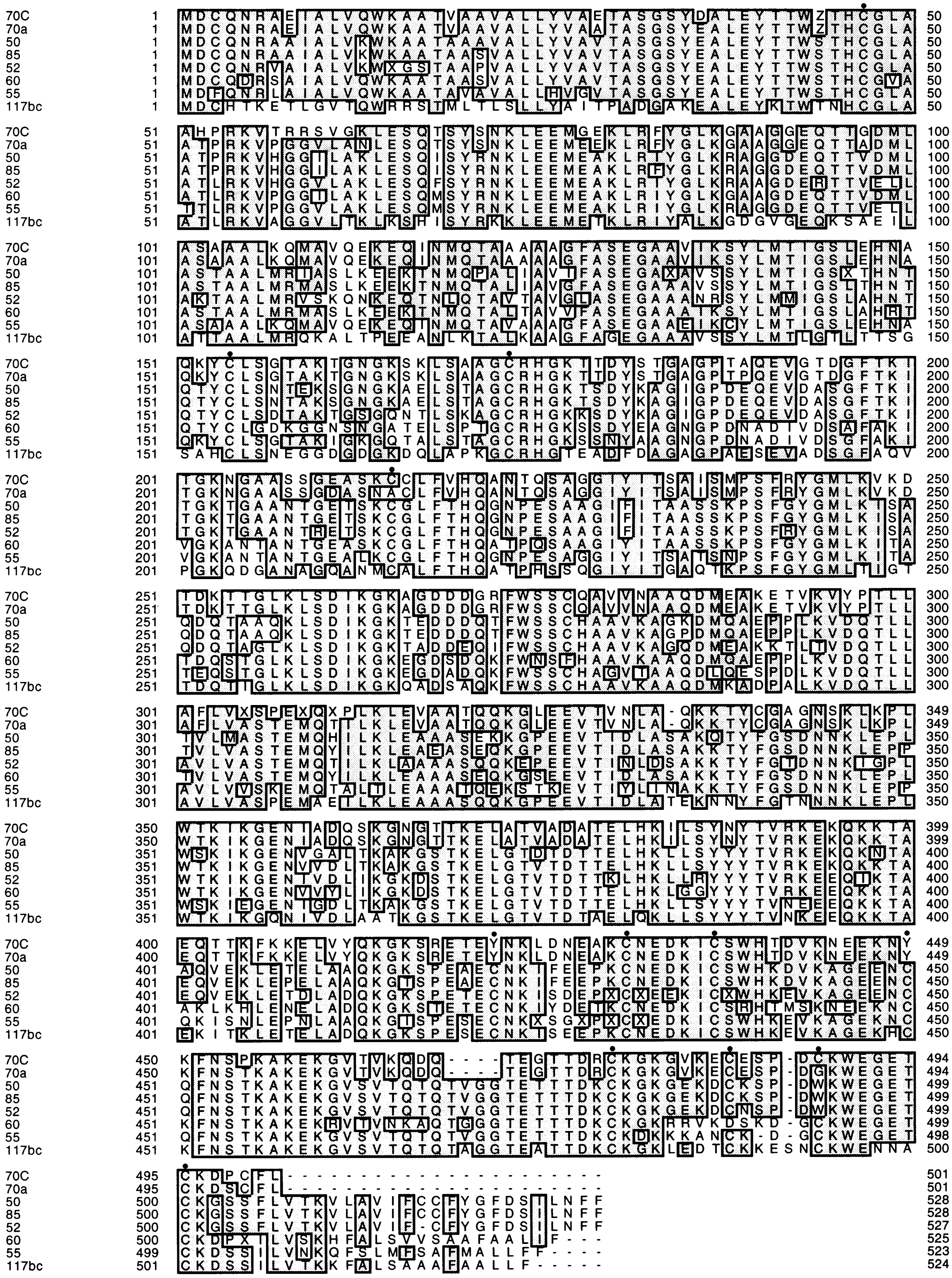
Taken together, analysis of the regions flanking the VSG ORF indicates the presence of URS sequences, both upstream and downstream (5′ to region 1 and in region 5), which are close to or include regions of high sequence diversity, while beyond this, similarity becomes greater (region 6 and the 3′ end of region 5). Clearly a role for the URS in the mobilization of VSG sequences is suggested by this, but the precise mechanism remains unknown.

Analysis of the Protein Coding Sequences

Following translation of the nucleotide sequences, the predicted amino acid sequences were aligned using Pileup (Fig. 5). The translation alignment shows the 117 basic copy to be more distantly related to the family members than the family members are to each other. This is exemplified both by the primary structure of the signal sequences, where the basic copy sequence is clearly very different from that of the family members (Fig. 5), and also from the phylogenetic analysis, which places the basic copy as an outlying molecular taxon (see below). In fact, the signal sequence is the most divergent region between the basic copy and the family member consensus. The positions of cysteines known to disulfide bond in VSG117 (Allen and Gurnett 1983) are highly conserved among the entire group (indicated by dots in Fig. 5), as reported previously for VSGs (discussed in Blum et al. 1993), consistent with all of the expressed proteins having similar secondary structures.

Each of the sequences encodes at least one potential N-glycosylation site. Except for pSUB70C all of the family members conserve the site at codon 453 of the basic copy. pSUB52 encodes one additional site at codon 136, and pSUB55 two new sites at codons 238 and 423. Both pSUB70A and pSUB70C encode two sites at 366 and 388. The possible contribution of an N-glycan to the antigenicity of the VSG is not known, although VSG 117 is known to use the site at codon 453 (Holder 1985).

We confirmed the presence of a TAA stop codon at amino acid 44 in pSUB70A (Beals and Boothroyd 1992b) and found the same feature in the closely related pSUB70C gene, presumably resulting from a single base change from the TCA serine codon in all of the other family members. We also observed that the reading frames of pSUB70A and pSUB70C are closed by a +1 frame-shift at codon 414 (data not shown). In order to allow comparison of the coding potential of the 70A and 70C genes we have artificially frame-shifted the nucleotide sequence by removing this adenine nucleotide for the translations shown in Fig. 5. Again, because regions of VSG genes can frequently be recombined in a segmental fashion, stop codons or frame-shifts do not preclude full functionality of the adjacent sequence. A re-



```
70C      1    MDCQNRAEIALVQWKAATVAAVALLYVAETASGSYDALEYTTWZTHCGLA   50
70a      1    MDCQNRAEIALVQWKAATVAAVALLYVAATASGSYEALEYTTWZTHCGLA   50
50       1    MDCQNRAAIALVKWKAATAAAVALLYVAVTASGSYEALEYTTWSTHCGLA   50
85       1    MDCQNRAAIALVKWKAATAASVALLYVAVTASGSYEALEYTTWSTHCGLA   50
52       1    MDCQNRVAIALVKWXGSTAAPVALLYVAVTASGSYEALEYTTWSTHCGLA   50
60       1    MDCQDRSAIALVQWKAATAASVALLYVAVTASGSYEALEYTTWSTHCGVA   50
55       1    MDFQNRLAIALVQWKAATAVAVALLHVGVTASGSYEALEYTTWSTHCGLA   50
117bc    1    MDCHTKETLGVTQWRRSTMLTLSLLYAITPADGAKEALEYKTWTNHCGLA   50

70C     51    AHPRKVTRRSVGKLESQTSYSNKLEEMGEKLRFYGLKGAAGGEQTTGDML  100
70a     51    ATPRKVPGGVLANLESQTSYSNKLEEMEEKLRFYGLKGAAGGEQTTADML  100
50      51    ATPRKVHGGILAKLESQISYRNKLEEMEAKLRTYGLKRAGGDEQTTVDML  100
85      51    ATPRKVHGGILAKLESQISYRNKLEEMEAKLRFYGLKRAGGDEQTTVDML  100
52      51    ATLRKVHGGVLAKLESQFSYRNKLEEMEAKLRTYGLKRAAGDERTTVELL  100
60      51    ATLRKVPGGILAKLESQMSYRNKLEEMEAKLRIYGLKGAAGDEQTTVDML  100
55      51    TTLRKVPGGVLAKLESQMSYRNKLEEMEAKLRIYGLKRAGGDEQTTVELL  100
117bc   51    ATLRKVAGGVLTKLKSHISYRKKLEEMETKLRIYALKGDGVGEQKSAEIL  100

70C    101    ASAAALKQMAVQEKEQINMQTAAAAAGFASEGAAVIKSYLMTIGSLEHNA  150
70a    101    ASAAALKQMAVQEKEQINMQTAAAAAGFASEGAAVIKSYLMTIGSLEHNA  150
50     101    ASTAALMRIASLKEEKTNMQPALIAVTFASEGAXAVSSYLMTIGSXTHNT  150
85     101    ASTAALMRMASLKEEKTNMQTALIAVGFASEGAAAVSSYLMTIGSLTHNT  150
52     101    AKTAALMRVSKQNKEQTNLQTAVTAVGLASEGAAANRSYLMMIGSLAHNT  150
60     101    ASTAALMRMASLKEEKTNMQTALTAVVFASEGAAATKSYLMTIGSLAHRT  150
55     101    ASAAALKQMAVQEKEQTNMQTAVAAAGFASEGAAEIKCYLMTIGSLEHNA  150
117bc  101    ATTAALMRQKALTPEEANLKTALKAAGFAGEGAAAVSSYLMTLGTLTTSG  150

70C    151    QKYCLSGTAKTGNGKSKLSAAGCRHGKTTDYSTGAGPTAQEVGTDGFTKI  200
70a    151    QKYCLSGTAKTGNGKSKLSAAGCRHGKTTDYSTGAGPTPQEVGTDGFTKI  200
50     151    QTYCLSNTEKSGNGKAELSTAGCRHGKTSDYKAGIGPDEQEVDASGFTKI  200
85     151    QTYCLSNTAKSGNGKAELSTAGCRHGKTSDYKAGIGPDEQEVDASGFTKI  200
52     151    QTYCLSDTAKTGSGQNTLSKAGCRHGKKSDYKAGIGPDEQEVDASGFTKI  200
60     151    QTYCLGDKGGNSNGATELSPTGCRHGKSSDYEAGNGPDNADIVDSAFAKI  200
55     151    QKYCLSGTAKIGKGQTALSTAGCRHGKSSNYAAGNGPDNADIVDSGFAKI  200
117bc  151    SAHCLSNEGGDGDGKDQLAPKGCRHGTEADFDAGAGPAESEVADSGFAQV  200

70C    201    TGKNGAASSGEASKCCLFVHQANTQSAGGTYITSAISMPSFRYGMLKVKD  250
70a    201    TGKNGAASSGDASNACLFVHQANTQSAGGIYITSAISMPSFRYGMLKVKD  250
50     201    TGKTGAANTGETSKCGLFTHQGNPESAAGIFITAASSKPSFGYGMLKTSA  250
85     201    TGKTGAANTGETSKCGLFTHQGNPESAAGIFITAASSKPSFGYGMLKISA  250
52     201    TGKTGAANTRETSKCGLFTHQGNPESAAGIFITAASSKPSFRYGMLKISA  250
60     201    VGKANTANTGEASKCGLFTHQATPQSAAGIYITAASSKPSFGYGMLKITA  250
55     201    VGKANTANTGEALKCGLFTHQGNPESAGGIYITSATSNPSFGYGMLKITA  250
117bc  201    PGKQDGANAGQANMCALFTHQATPHSSQGIYITGAQTKPSFGYGMLTIGT  250

70C    251    TDKTTGLKLSDIKGKAGDDDGRFWSSCQAVVNAAQDMEAKETVKVYPTLL  300
70a    251    TDKTTGLKLSDIKGKAGDDDGRFWSSCQAVVNAAQDMEAKETVKVYPTLL  300
50     251    QDQTAAQKLSDIKGKTEDDDQTFWSSCHAAVKAGKDMQAEPPLKVDQTLL  300
85     251    QDQTAAQKLSDIKGKTEDDDQTFWSSCHAAVKAGPDMQAEPPLKVDQTLL  300
52     251    QDQTAGLKLSDIKGKTADDEQIFWSSCHAAVKAGQDMEAKKTLTVDQTLL  300
60     251    TDQSTGLKLSDIKGKEGDSDQKFWNSFHAAVKAAQDMQAEPPLKVDQTLL  300
55     251    TEQSTGLKLSDIKGKEGDADQKFWSSCHAGVTAAQDLQESPDLKVDQTLL  300
117bc  251    TDQTIGLKLSDIKGKQADSAQKFWSSCHAAVKAAQDMKADPALKVDQTLL  300

70C    301    AFLVXSPEXQXPLKLEVAATQQKGLEEVTVNLA-QKKTYCGAGNSKLKPL  349
70a    301    AFLVASTEMQTTLKLEVAATQQKGLEEVTVNLA-QKKTYCGAGNSKLKPL  349
50     301    TVLMASTEMQHILKLEAAASEKKGPEEVTIDLASAKQTYFGSDNNKLEPL  350
85     301    TVLVASTEMQYILKLEAEASEQKGPEEVTIDLASAKKTYFGSDNNKLEPP  350
52     301    AVLVASTEMQTILKLAAAASQQKEPEEVTINLDSAKKTYFGTDNNKIGPL  350
60     301    TVLVASTEMQYILKLEAAASEQKGSEEVTIDLASAKKTYFGSDNNKLEPL  350
55     301    AVLVVSKEMQTALTLEAAATQEKSTKEVTIYLINAKKTYFGSDNNKLEPP  350
117bc  301    AVLVASPEMAETLKLEAAASQQKGPEEVTIDLATEKNNYFGTNNNKLEPL  350

70C    350    WTKIKGENIADQSKGNGTTKELATVADATELHKILSYNYTVRKEKQKKTA  399
70a    350    WTKIKGENIADQSKGNGTTKELATVADATELHKILSYNYTVRKEKQKKTA  399
50     351    WSKIKGENVGALTKAKGSTKELGTDTDTTELHKLLSYYYTVRKEKQKNTA  400
85     351    WTKIKGENVVDLTKAKGSTKELGTVTDTTELHKLLSYYYTVRKEKQKKTA  400
52     351    WTKIKGENTVDLIKGKDSTKELGTVTDTTKLHKLLRYYYTVRKEEQIKTA  400
60     351    WTKIKGENVVYLIKGKDSTKELGTVTDTTELHKLGGYYYTVRKEEQKKTA  400
55     351    WSKIEGENTGDLTKAKGSTKELGTVTDTTELHKLLSYYYTVNEEEQKKTA  400
117bc  351    WTKIKGQNIVDLAATKGSTKELGTVTDTAELQKLLSYYYTVNKEEQKKTA  400

70C    400    EQTTKFKKELVYQKGKSRETEYNKLDNEAKCNEDKICSWHTDVKNEEKNY  449
70a    400    EQTTKFKKELVYQKGKSRETEYNKLDNEAKCNEDKICSWHTDVKNEEKNY  449
50     401    AQVEKLETELAAQKGKSPEAECNKIFEEPKCNEDKICSWHKDVKAGEENC  450
85     401    EQVEKLEPELAAQKGTSPEAECNKIFEEPKCNEDKICSWHKDVKAGEENC  450
52     401    EQVEKLETDLADQKGKSPETECNKISDEPXCXEEKICXWHKEVKAGEENC  450
60     401    AKLKHLENELADQKGKSTETECNKIYDETKCNEDKICSRHTMSKNEEKNC  450
55     401    QKISNLEPNLAAQKGISPESECNKXSGXPXCXEDKICSWHKEVKAGEKNC  450
117bc  401    EKITKLETELADQKGKSPESECNKTSEEPKCNEDKICSWHKEVKAGEKHC  450

70C    450    KFNSPKAKEKGVTVKQDQ----TEGTTDRCKGKGVKECESP-DCKWEGET  494
70a    450    KFNSTKAKEKGVTVKQDQ----TEGTTDRCKGKGVKECESP-DGKWEGET  494
50     451    QFNSTKAKEKGVSVTQTQTVGGTETTTDKCKGKGEKDCKSP-DWKWEGET  499
85     451    QFNSTKAKEKGVSVTQTQTVGGTETTTDKCKGKGEKDCKSP-DWKWEGET  499
52     451    QFNSTKAKEKGVSVTQTQTVGGTETTTDKCKGKGEKDCNSP-DWKWEGET  499
60     451    KFNSTKAKEKRVTVNKAQTGGGTETTTDKCKGKRRVKDSKD-GCKWEGET  499
55     451    QFNSTKAKEKGVSVTQTQTVGGTETTTDKCKDKKKANCK-D-GCKWEGET  498
117bc  451    KFNSTKAKEKGVSVTQTQTAGGTEATTDKCKGKLEDTCKKESNCKWENNA  500

70C    495    CKDPCFL----------------------                        501
70a    495    CKDSCFL----------------------                        501
50     500    CKGSSFLVTKVLAVIFCCFYGFDSILNFF                        528
85     500    CKGSSFLVTKVLAVIFCCFYGFDSILNFF                        528
52     500    CKGSSFLVTKVLAVIF-CFYGFDSILNFF                        527
60     500    CKDPXILVSKHFALSVVSAAFAALIF---                        525
55     499    CKDSSILVNKQFSLMFSAFMALLFF----                        523
117bc  501    CKDSSILVTKKFALSAAAFAALLF-----                        524
```

Fig. 5. Translation of the coding region of the DNA sequences of the 117 VSG family. Sequences are numbered from the initiation methionine, and the first amino acid of the mature protein is residue 34 by analogy to the known mature N-terminus of VSG 117. Predicted amino acids are given in the *single-letter code.* Unassigned amino acids are designated by an *X* and spaces introduced into the alignment by a ‘‘−’’.*Z* indicates in frame stop codons. Cysteines known to disulfide bond in VSG117 are indicated by a *dot* above the alignment. Identities in four or more sequences at a given position are *shaded.* Note that a frameshift in both pSUB70A and pSUB70C at position 414 is ignored (see text).

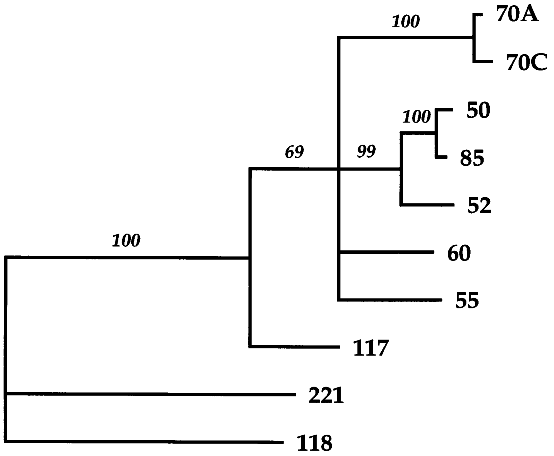


Fig. 6. Molecular phylogenetic reconstruction for the 117 VSG gene family. The topology was calculated using PAUP in branch and bound mode. The translated sequences of the 118 and 221 VSGs were included as an outgroup. The resulting tree was then subjected to robustness analysis by bootstrapping (10^3 replicates) and the resultant final topology is shown, together with the bootstrap values. Note that the tree places the 117 basic copy as an outlying node, with all the family members derived from a single branch. The relative positions of the nodes within the family member branch cannot be determined with confidence, and therefore PAUP shows them as radiating from a single node.

cent report documents a further example of a VSG basic copy which contains an in-frame stop codon (Aline et al. 1994).

The rationale for retention of such VSG sequences may be that the simple antigenic repertoire may be exhausted after many switches, and therefore the production of novel combinations of sequence may be a strategy to allow continued evasion of the host immune system. Such recombinations may be rare, and therefore not readily observed early on in infection, but as the immunological pressure to generate novel epitopes becomes more acute these events may show a greater probability of being represented as a surviving clone as seen among late-expressing variants (Longacre and Eisen 1986).

With the exception of pSUB70A, pSUB70C, and pSUB60 all of the sequences predict functional GPI-single peptides, and the GPI-attachment site XaaSerSer is conserved. In the case of pSUB70A and pSUB70C, the ORF is truncated at this point, removing the hydrophobic C-terminal peptide as well as most of the hydrophilic spacer. pSUB60 has a sufficiently long C-terminal hydrophobic peptide, but the presence of an unassigned amino acid at position 504 precludes judgment of the potential functionality of this sequence. Since activation of VSG genes typically includes fusion to the existing 3′ region of an ELC (Bernards et al. 1981; Thon et al. 1989, 1990), including the GPI-signal peptide region, the absence of this sequence in some of the family member genes in no way precludes their being fully functional in antigenic variation.

The 117 VSG ORFs were analyzed using PAUP (Swafford 1993) in exhaustive mode to determine a probable evolutionary relationship between the eight genes. The sequences of the ORFs of 118 and 221 VSGs were included to provide an outgroup. The single derived tree obtained was subjected to a bootstrap branch and bound algorithm (Swafford 1993) to test the robustness of the topology, and the bootstrap result is shown (Fig. 6). The bootstrap values are very high for all of the internodes except that containing all of the family members; i.e., it is not possible to assign a relative divergence order for the pSUB70A and pSUB70C clade; the pSUB50, pSUB52, and pSUB85 clade; or pSUB60 and pSUB55. This topology shows the 117 basic copy to be an outlier compared with the remainder of the family and is consistent with a duplication of the 117 basic copy and subsequent radiation of the family members from this ancestor. It is also possible that the divergence of the basic copy from the other 117 VSG genes has been accelerated by virtue of more frequent selection for variation by exposure to the host immune system as only the 117 BC is definitively known to be expressed.

The family members include two clusters, with pSUB70A and pSUB70C present on one node, and pSUB50, pSUB52, and pSUB85 on a second. The relationships are consistent with a steady duplication and divergence of the 117 VSG genes with time following an initial duplication (see above). Interestingly, the accumulation of mutations within the ORF is biased against insertion/deletion, and nucleotide replacement is favored. The close clustering of pSUB70A and pSUB70C suggests that the duplication of these genes is a recent event. Therefore it is most probable that the acquisition of the in-frame stop codon and the frame-shift mutation predate the duplication event, while the distance of these two genes from the rest of the cluster suggests that the common ancestor of this pair has been diverging either for a long period of time or divergence has been accelerated due to the loss of function.

The aligned sequences were analyzed for regions of reduced and enhanced variability using PlotSimilarity (Fig. 7). When all of the 117 VSG gene family, including the basic copy, were considered, two strong and one weaker peak of increased diversity were detected in the region corresponding to the mature VSG. The high level of variability at the extreme C-terminus is due to truncation of the GPI-signal in pSUB70A and pSUB70C and reflects the extreme diversity seen at the nucleotide level in this region (region 4, see above). All three variable peaks were within the N-terminal domain, and had approximate maxima at amino acids 115 (peak I), 155 (peak II), and 185 (peak III). Peaks I and II encompass ~10–15 amino acids, while peak III is broader and less pronounced.

The N-termini are the most variable among VSGs, presumably because they are the portions of the molecule

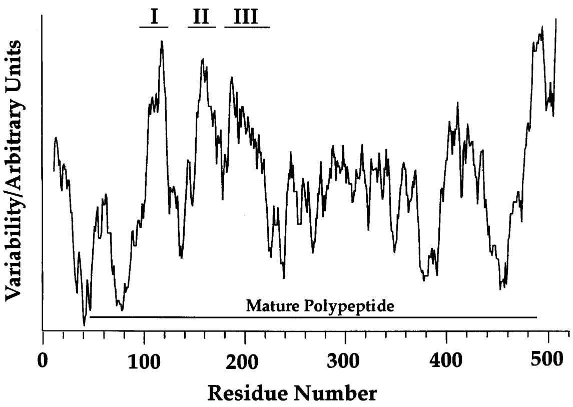


Fig. 7. Localization of Variability in the 117 VSG family coding sequences. Hypervariability of the translated 117 family coding sequences is seen in three discrete N-terminal regions. The sequences shown in Fig. 5 were compared using PlotSimilarity with a window of 20 residues. The three hypervariable peaks within the N-terminal domain are indicated by *I, II,* and *III.* The high variability score at the C-terminus is due to the different lengths of the GPI-signal peptides. The extent of the mature polypeptide, based on analysis of the 117 VSG protein, is indicated.

that are exposed to the host environment. However, there are considerable constraints on the amount of variability that can be tolerated, and it appears that different VSGs may adopt similar tertiary structures (Olafson et al. 1984; Blum et al. 1993). When the sequences 50, 52, and 85 alone were considered, only peaks 1 and II were detected (data not shown). These three VSGs are clustered in the PAUP analysis, i.e., are most closely related, suggesting that regions I and II are evolving the most rapidly, while variation within peak III is only observed when more distant relationships are considered, suggesting less rapid evolution.

Inspection of the sequences encoding the hypervariable peaks with the alignments presented by Blum et al. (1993), together with the two available X-ray crystal structures (Blum et al. 1993; Freymann et al. 1990), indicates that peaks II and III are positioned within the solvent accessible portion of the molecule, in the surface-exposed loops, whereas peak I is within the long helix ‘‘B’’ that is buried within the folded VSG. The localization of variable regions II and III to surface-exposed loops is clearly consistent with the selection of new variants by virtue of their ability to evade the host immune response by presentation of novel surface epitopes. Our assumption that the 117 VSG family will adopt a similar fold to the VSGs of MITat 1.2 (VSG 221) and ILTat 1.24 is not unreasonable as all three are type-A VSGs based on sequence comparisons (as defined in Blum et al. 1993) and the latter two are no more similar to each other than either is to VSG 117.

The presence of a variable region within the conserved helix B is unexpected. It could indicate that this region is under pressure to change, perhaps because this region impacts the B-cell epitopes, as suggested by epitope mapping studies (Hsia et al. 1996) or because the helix comprises a T-cell epitope. Alternatively, it could be reflecting background variation in a region where variance is tolerated provided that it does not interfere with the α-helical structure.

There are two regions of increased similarity within the C-terminus at amino acid positions 380 and 460. As the C-terminus of VSG has not been crystallized it is not possible to determine if this decrease in variability corresponds to a specific structural feature. Secondary-structure prediction algorithms did not point out any remarkable features about this region, but the possibility of selection for retention of a specific sequence may indicate some structural or functional importance in this region. For example, the minimum in variation at amino acids 450–460 corresponds to the location of a conserved glycosylation site at Asn 452/3 in all of the 117 VSG genes except pSUB70C (Fig. 5). If addition of carbohydrate at this location is functionally important the polypeptide sequence surrounding the site may also be constrained because the secondary structure of the N-glycosylation site may influence its ability to be processed (Kornfeld and Kornfeld 1985).

Our results stand in marked contrast with what has been seen for other parasite surface antigens where more clear, functional constraints are operating; for example, the gp63 gene family of Leishmania is less diverse and its variation is far more localized, probably because of the need to retain metalloprotease activity (Roberts et al. 1993). The data presented here also indicate that the total VSG repertoire is hierarchical at many levels, but most importantly in genomic context, influencing the probability and mechanism of activation, and in coding potential, affecting the amount of diversity once expressed. Finally, the information presented here gives new hints of some of the pressures (e.g., T-cell selection) and mechanics (e.g., the recombinogenic/mutagenic repeats) that may accelerate the evolution of this remarkable system for antigenic variation.

Acknowledgments. Our thanks are due to George Yeltzine (PAN Facility, Stanford University) for 373 DNA sequencer work; to Helen Field (Applied Biosystems Inc.) for sequence determination of the true 117 basic copy and distant upstream regions and for advice on template purification; to Adrian Lawrence, Ian Manger, and Keith Wilson (Stanford University) for overall help, especially with positioning region 6 and alignment analyses; to Thomas Beals (University of California at Los Angeles) for providing reagents and unpublished sequence data; to Richard Davis (San Francisco State University) for assistance with GeneWorks; and to Keith Wilson and Thomas Beals for comments on the manuscript.

References

Allen G, Gurnett L (1983) Location of the six disulphide bonds in a variant surface glycoprotein (VSG117) from *Trypanosoma brucei.* Biochem J 209:481–487

Aline RF, Myler PJ, Gobright E, Sturat KD (1994) Early expression of a *Trypanosoma brucei* VSG gene duplicated from and imcomplete basic copy. J Euk Microbiol 41:71–78

Barbet AF, Kamper SM (1993) Importance of mosaic gene expression to survival of *Trypanosoma brucei.* Parasitol Today 9:63–66

Beals TP, Boothroyd JC (1992a) Genomic organization and context of a trypanosome variant surface glycoprotein gene family. J Mol Biol 225:961–971

Beals TP, Boothroyd JC (1992b) Sequence divergence among members of a trypanosome variant surface glycoprotein gene family. J Mol Biol 225:973–983

Bernards A, Van der Ploeg LHT, Frasch ACC, Borst P, Bothroyd JC, Coleman S, Cross GAM (1981) Activation of trypanosome surface glycoprotein genes involves a gene duplication-transposition leading to an altered 3′ end. Cell 27:497–505

Blum ML, Down JA, Gurnett AM, Carrington M, Turner MJ, Wiley DC (1993) A structural motif in the variant surface glycoproteins of *Trypanosoma brucei.* Nature 362:603–609

Boothroyd JC, Cross GAM (1982) Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5′ end. Gene 20:281–289

Boothroyd JC, Paynter CA, Coleman SL, Cross GAM (1982) Complete nucleotide sequence of cDNA coding for a variant surface glycoprotein of *Trypansoma brucei,* J Mol Biol 157:547–556

Borst P, Cross GAM (1982) Molecular basis for trypanosome antigenic variation. Cell 29:291–303

Campbell DA, van Bree MP, Boothroyd JC (1984) The 5′-limit of transposition and upstream barren region of a trypanosome VSG gene: tandem 76 base-pair repeats flanking $(TAA)_{90}$. Nucleic Acids Res 12:2759–2774

Cross GAM (1990a) Cellular and genetic aspects of antigenic variation in trypanosomes. Annu Rev Immunol 8:83–110

Cross GAM (1990b) Glycolipid anchoring of plasma membrane proteins. Annu Rev Cell Biol 6:1–39

Devereux J, Haeberli P, Smithies O (1984) A set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Freymann D, Down J, Carrington M, Roditi I, Turner M, Wiley DC (1990) A resolution structure of the N-terminal domain of a variant surface glycoprotein from *Trypanosoma brucei.* J Mol Biol 216: 141–160

Hasan G, Turner MJ, Cordingley JS (1984) Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei.* Cell 37:333–341

Holder AA (1985) Glycosylation of the variant surface antigens of *Trypanosoma brucei.* Curr Top Microbiol Immunol 117:57–74

Hsia R-C, Beals TP, Boothroyd JC (1996) Use of chimeric recombinant polypeptides to analyze conformational, surface epitopes on trypanosome variant surface glycoproteins. Mol Microbiol 19:53–63

Kamper SM, Barbet AF (1992) Surface epitope variation via mosaic gene formation is potential key to long-term survival of *Trypanosoma brucei.* Mol Biochem Parasitol 53:33–44

Kimmel BE, Ole-Moiyoi OK, Young JR (1987) Ingi, a 52-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol Cell Biol 7:1465–1475

Kornfeld R, Kornfeld S (1985) Assembly of asparagine-linked oligosaccharides. Annu Rev Biochem 54:631–664

Lamont GS, Tucker RS, Cross GAM (1986) Analysis of antigen switching rates in *Trypanosoma brucei.* Parasitology 92:355–367

Liu AYC, Van der Ploeg LHT, Rijsewijk FAM, Borst P (1983) The transposition unit of VSG gene 118 of *Trypanosoma brucei:* presence of repeated elements at its border and absence of promoter associated sequences. J Mol Biol 167:57–75

Longacre S, Eisen H (1986) Expression of whole and hybrid genes in *Trypanosoma equiperdum* antigenic variation. EMBO J 5:1057–1063

Lu Y, Hall T, Gay LS, Donelson JE (1993) Point mutations are associated with a gene duplication leading to the bloodstream reexpression of a trypanosome metacyclic VSG. Cell 72:397–406

Maniatis T, Sambrook J, Fritsch J (1985) Molecular cloning: a laboratory manual, 1st ed. Cold Spring Harbor Press, Cold Spring Harbor, NY

Olafson RW, Clarke MW, Kieland SL, Pearson TW, Barbet AF, McGuire T (1984) Amino terminal sequence homology among variant surface glycoproteins of African trypanosomes. Mol Biochem Parasitol 12:287–298

Pays E, Vanhamme L, Berberof M (1994) Genetic controls for the expression of surface antigens in African trypanosomes. Annu Rev Micro 48:25–52

Roberts SC, Swihart KG, Agey MW, Ramamoorthy R, Wilson ME, Donelson JE (1993) Sequence diversity and organization of the msp gene family encoding gp63 of *Leishmania chagasi.* Mol Biochem Parasitol 62:157–172

Smiley BL, Aline RF, Myler PJ, Stuart K (1990) A retroposon in the 5′ flank of a *Trypanosoma brucei* VSG gene lacks insertional terminal repeats. Mol Biochem Parasitol 42:143–152

Swafford DL (1993) Phylogenetic Analysis Using Parsimony, Version 311. Illinois History Survey, Champaign, IL

Thon G, Baltz T, Eisen H (1989) Antigenic diversity by the recombination of pseudogenes. Genes Dev 3:1247–1254

Thon G, Baltz T, Giroud C, Eisen H (1990) Trypanosome variable surface glycoproteins: composite genes and order of expression. Genes Dev 9:1374–1383

Van der Ploeg LHT (1990) In: Hames BD, Glover DM (eds) Genome rearrangement. Oxford University Press, Oxford

Van der Ploeg LHT, Valerio D, De Lange T, Bernards A, Borst P, Grosveld FG (1982) An analysis of cosmid clones of nuclear DNA from *Trypanosoma brucei* shows that the genes for variant surface glycoproteins are clustered in the genome. Nucleic Acids Res 10: 5905–5923

Vickerman K (1989) Trypanosome sociology and antigenic variation. Parasitology 99:537–547