Reprinted from *J. Mol. Biol.* (1991) **221**, 823–835

Variant Specific Glycoprotein of *Trypanosoma brucei* Consists of Two Domains each Having an Independently Conserved Pattern of Cysteine Residues

Mark Carrington, Nancy Miller, Michael Blum, Isabel Roditi, Don Wiley and Mervyn Turner

J. Mol. Biol. (1991) **221**, 823–835

Variant Specific Glycoprotein of *Trypanosoma brucei* Consists of Two Domains each Having an Independently Conserved Pattern of Cysteine Residues

Mark Carrington[1]†, Nancy Miller[1], Michael Blum[2], Isabel Roditi[3] Don Wiley[2] and Mervyn Turner[4]

[1]*Department of Biochemistry, Tennis Court Road Cambridge, CB2 1QW, U.K.*

[2]*Department of Biochemistry & Molecular Biology Harvard University, 7 Divinity Avenue Cambridge, MA 01238, U.S.A.*

[3]*Institut für Allgemeine Mikrobiologie, Universität Bern Baltzerst. 4, CH 3012 Bern, Switzerland*

[4]*Merck, Sharp & Dohme Research Laboratories PO Box 2000, Rahway, NJ 07065, U.S.A.*

(Received 27 November 1990; accepted 19 June 1991)

The complete amino acid sequences for nine variant specific glycoproteins (VSGs) of *Trypanosoma brucei* are presented. These have more than doubled the size of the VSG sequence data base and have enabled a new and more rigorous comparison to be made between amino acid sequences of different VSGs. Each VSG can be defined as a combination of an N-terminal domain type and a C-terminal domain type, based on the distribution of cysteine residues within the molecule. This identifies three N-terminal domain types and at least four C-terminal domain types. Different combinations of N and C-terminal domains can be formed; for example, in the sequences presented here, two different N-terminal domains are found in association with each of three different C-terminal domains. The biological context of the domain structure of VSGs is discussed.

Keywords: *Trypanosoma brucei*; VSG, sequence comparison; domains; cysteine conservation

1. Introduction

A population of the protozoan parasite *Trypanosoma brucei* can persist in the bloodstream of the mammalian host through the process of antigenic variation. The surface of the bloodstream-form of the parasite is covered by a 12 to 15 nm thick coat (Vickerman, 1969) composed of a single polypeptide species, the variant specific glycoprotein (VSG‡; Cross, 1975). Antigenic variation occurs when an individual trypanosome stops expressing one VSG gene and starts expressing another which encodes an antigenically distinct VSG (for reviews, see Pays & Steinert, 1988; Borst, 1986) thus delaying clearance by the host immune system. Direct estimates have indicated that there are approximately 1000 VSG genes in the trypanosome genome (Van der Ploeg *et al.*, 1982) and although no estimate of the number of functional, antigenically distinct, VSGs that can be generated from these is available, infection of a rabbit with a single cell from the related species *Trypanosoma equiperdum* led to the appearance of more than 100 serologically distinct variants (Capbern *et al.*, 1977).

Biochemical and sequence analysis of VSGs has shown that the mature polypeptides contain between 400 and 500 residues and have at least one *N*-linked oligosaccharide, although the position of this oligosaccharide varies (Cross, 1984; Turner, 1988). VSGs are found as dimers in solution (Auffret & Turner, 1981) and are believed to exist as dimers or higher oligomers at the cell surface (Strickler & Patton, 1982). Attachment to the cell surface is

† Author to whom all correspondence should be addressed.

‡ Abbreviations used: VSG, variant specific glycoprotein; PCR, polymerase chain reaction.

0022–2836/91/190823–13 $03.00/0

through a glycosyl–phosphatidylinositol anchor linked to the C-terminal amino acid of the mature polypeptide (Cardoso de Almeida & Turner, 1983; Ferguson *et al.*, 1988). An N-terminal domain can be physically separated through the action of exogenous proteases (Johnson & Cross, 1979). This domain contains two-thirds to three-quarters of the amino acid sequence and also exists as a dimer in solution. The N-terminal domain has now been crystallised for several VSGs (Metcalf *et al.*, 1988) and the crystal structure has been solved at high resolution for two of them (Freymann *et al.*, 1984, 1990; Metcalf *et al.*, 1987; M. Blum *et al.*, unpublished results). The C-terminal domain contains the balance of the sequence. At present, the structure for neither an intact VSG nor an isolated C-terminal domain has been reported.

The complete sequence of seven *T. brucei* VSGs have been published, derived either directly from the polypeptide (MITat 1·4; Allen *et al.*, 1982); and/or from the sequence of cDNA (MITat 1·4, Boothroyd *et al.*, 1982; ANTat 1·1, Michiels *et al.*, 1983, Pays *et al.*, 1983; ANTat 1·10, Pays *et al.*, 1983; MVAT 4, MVAT 7, Leonardo *et al.*, 1984; ILTat 1·3, Rice-Ficht *et al.*, 1981) or a combination of cDNA and genomic clones (ILTat 1·1, Rice-Ficht *et al.*, 1982). It should be noted that (1) ANTat 1·1 and ANTat 1·10 have >75% amino acid sequence identity and are encoded by members of the same gene family despite not cross-reacting antigenically when expressed on living trypanosomes and (2) a partial polypeptide sequence, derived from cDNA sequence, is available for ANTat 1·8 (Matthyssens *et al.*, 1981), this has 104/115 identities with MITat 1.4 and so both are probably members of the same VSG gene family.

Comparison of these sequences has shown that the sequence of the N-terminal domain is extremely diverse except for the relative locations of cysteine residues which are conserved (Cross, 1984). The C-terminal domain exhibits a degree of sequence conservation (Rice-Ficht *et al.*, 1981, Holder & Cross, 1981, Englund *et al.*, 1988). This homology is found both near the C terminus of the mature polypeptide, and in a hydrophobic C-terminal extension present in a VSG precursor, detected by comparison of cDNA and polypeptide sequence data (Holder & Cross, 1981), that presumably acts transiently as a membrane anchor before glycolipid addition. As in the N-terminal domain the conservation of cysteine residues is particularly strong in the C-terminal domain of the mature polypeptide.

In this paper a further nine VSG sequences are presented. These additional data have enabled a more coherent analysis of conservation in the primary sequences of VSGs. Each VSG can be described as a combination of an N-terminal domain type and a C-terminal domain type based on the distribution of cysteine residues. The N-terminal domain types and C-terminal domain types are interchangable so that, in different VSGs, one N-terminal domain type can be found in combination with most C-terminal domain types and *vice versa*.

The domain structures for nine members of the ILTAR 1 serodeme (Miller & Turner, 1981) are therefore now known, and the relationship between domain structure and switching can now be described for this serodeme.

2. Materials and Methods

(a) Trypanosoma brucei *stocks*

All trypanosomes used originated from the MITAR 1 (Cross & Manning, 1973) or ILTAR 1 (Miller & Turner, 1981) serodemes.

(b) *Oligodeoxynucleotides*

Oligonucleotides were synthesized using a Biosearch cyclone DNA synthesizer. Oligonucleotide 39 (5′ GTGTTAAAATATATCA 3′) was designed to base-pair with a conserved-sequence nucleotide sequence found in the 3′ untranslated region of VSG mRNAs (Majumder *et al.*, 1981). Oligonucleotide 41 (5′ GTTTCTGTACTATATTG 3′) was designed to base-pair with the reverse complement of the 3′ part of the mini-exon found at the 5′ end of all *T. brucei* mRNAs (Cornelissen *et al.*, 1986).

(c) *Preparation and sequencing of VSG cDNAs*

Polyadenylated RNA was purified from bloodstream trypanosomes (Roditi *et al.*, 1987). One of two procedures was then followed; either a cDNA library was constructed in λ NM1149 (Roditi *et al.*, 1987) or double stranded VSG cDNA was prepared using primer extension with oligonucleotide 39 followed by the polymerase chain reaction (PCR) using oligos 39 and 41 (Kawasaki, 1990). Clones from cDNA libraries containing cDNA derived from VSG mRNA, were identified using [^{32}P]oligonucleotide 39 and, if necessary, further clones isolated by screening with purified cDNA from initial isolates. PCR products were cloned into a plasmid vector. All cDNAs were sequenced by generating random subclones by sonication (Baer *et al.* 1984), which were cloned into M13 mp18 and sequenced by the dideoxy method (Sanger *et al.*, 1977).

3. Results

(a) *Synthesis and sequencing of cDNAs encoding VSGs*

The cDNA sequences are contained in the EMBL database, accession numbers are shown by each amino acid sequence in Figure 1. With the exception of MITat 1·2 these cDNA sequences contain part or all of the *T. brucei* mini-exon at their 5′ ends. In the case of the MITat 1·2 the first 16 bases of the sequence are taken from previously published data (Boothroyd *et al.*, 1981). From results using primer extension there are approximately 45 bases 5′ of the sequence presented before the mini-exon (data not shown). At or near the 3′ end of each cDNA sequence is the conserved region found near the 3′ end of VSG mRNAs (Majumder *et al.*, 1981). Taken together these two sequence blocks indicate that the data presented represents complete VSGs. The origin of the cDNAs sequenced were as follows: MITat 1·2, ILTat 1·24 and ILTat 1·25, cDNAs

MITat 1·1 (MIAG 060) EMBL X56761

MATGRAKNTK WARWLSTAGL IIVVTLPATT MAAERTGLKA TAWKPLCKLT TELSKVSGEM 60

LNKGQEVISN IQKIKAAEYK VSIYLAKNPE TQALQQSTLL RGYFARKTNG GLESYKTMGL 120

ATQIRSARAA AYLKGSIDEF LNLLESLKGG SENKCLVTTN ADTAATRRET KLDDQECALS 180

MPETKPEAAT RTELTQTGYP NLQHGGGGTA NTFQPTTSTG TCKLLSGHST NGYPTTSALD 240

TTAKVLAGYM TIPNTQVEAT LANMQAMGNG HKATAPAWHE AWEARNREAK AKDLAYTNET 300

GNLDTQPTLK ALVKTLLLPK DNTEHNAEAT KLEALFGGLA ADKTKTYLDM VDAEIIPAGI 360

AGRTTEAPLG KIHDTVELGD ILSNYEMIAA QNVVTLKKNL DAVSKKQQTE SAENKEKICN 420

AAKDNQKACE NLKEKGCVFN TESNKCELKK DVKEKLEKES KETEGKDEKA NTTGSNSFLI 480

HKAPLLLAFL LF

MITat 1·2 (MIAG 221) EMBL X56762

MPSNQEARLF LAVLVLAQVL PILVDSAAEK GFKQAFWQPL CQVSEELDDQ PKGALFTLQA 60

AASKIQKMRD AALRASIYAE INHGTNRAKA AVIVANHYAM KADSGLEALK QTLSSQEVTA 120

TATASYLKGR IDEYLNLLLQ TKESGTSGCM MDTSGTNTVT KAGGTIGGVP CKLQLSPIQP 180

KRPAATYLGK AGYVGLTRQA DAANNFHDND AECRLASGHN TNGLGKSGQL SAAVTMAAGY 240

VTVANSQTAV TVQALDALQE ASGAAHQPWI DAWKAKKALT GAETAEFRNE TAGIAGKTGV 300

TKLVEEALLK KKDSEASEIQ TELKKYFSGH ENEQWTAIEK LISEQPVAQN LVGDNQPTKL 360

GELEGNAKLT TILAYYRMET AGKFEVLTQK HKPAESQQQA AETEGSCNKK DQNECKSPCK 420

WHNDAENKKC TLDKEEAKKV ADETAKDGKT GNTNTTGSSN SFVISKTPLW LAVLLF

MITat 1·5 (MIAG 118) EMBL X56763

MIHSNKVATV VLALISSWPA DGTNNHGLKL QKAQAICKMS KELKATAMRA ANDAKLKITE 60

ILELENVFAA MIPNATKGTE ADGCTDYNAV FLEANNTAAE TVSKIATLAE SATKAAGAAG 120

RAAGVLDEFI AALAQAQGAT GLYCIQGSGT GAATHTELAD CFNGDLKPRN MLSIRDPKVS 180

AAATGSTDLT TLAKAMAASG TDTTFHGDQQ SKGCGLMKGT SDGIMIGQAL TGTFAWAQGL 240

LRFGALGANG IASTGVTGYA HTATASGNGV HWASDPEKIP VIAEAIALVS NYNTLADSIG 300

TRAKDAIEKV KKCMKATNKE IKREHIFLNV SHLNRELQKA VTELDKALNK QDAKAEAKQQ 360

PNCDDKKQIE CGDTPGCGWH KAEGKCEAKD GEGQKNQATG EKDANKNRCT QHGTNKEACE 420

KENTPGQSAV CGFRKGKDGE TDEPDKEKCR NGSFLTSKQF AFSVVSAAFM ALLF

Figure 1. The amino acid sequences of VSGs from the MITAR 1 and ILTAR 1 serodeme. Each sequence is identified by the encoded VSG and the MIAG clone number of the trypanosome population and then the accession number of the cDNA sequence in the EMBL database. The N-terminal signal sequence and the C-terminal hydrophobic extension absent in the mature polypeptide are underlined.

MITat 1·6 (MIAG 151) EMBL X56764

MAVHRALAAY AISLYVLLPR KSGATDKGAI KFETWEPLCL LTQDFGNLYN RAHKLNLDID 60

TYVTAAQADQ LRLQILLSRA SSKIEAAAAA AATAAIAADL AGKAKHVASC KLAATTLTAT 120

TGYLHGRIAE FLEVMTAHRG TTNKYGCLSK SRSDNSNSIT DSVANLKDKC KLTVQQISPD 180

NKATEQITKA GFTKLTATGG LTSSDLGGSG QAVCMILSTT ASEVVNNGNL DEPVPYAGGY 240

LRRKHDLTSD GNDNLATITD SASAPATRAK TDPYLQIWRA FKNLEDCEST FTSGYSRPSP 300

ETLKAADETK TAIKNYVVQK EGKYDQATDK EDDYKNLDKI FKDGKDFYPQ KLWDAMDKKD 360

LLKDATQTNE IKKLADITDR SELNKVLLYY TRQKEQTLTK ELKEAQEKAT QANQNDAAAK 420

AAEDSCNKLV GGEKCNADKK CSYETETDGT KKCKFNATKA EKSGAPVTQA QTVGETEATP 480

EKCKGKDAKT CGTTQGCKWE GETCKDSSIL VTKKFALTVV SAAFVALLF

ILTat 1·21 (MIAG 201) EMBL X56766

MLRALLPSTT LALILAGGGH AAVGDAFPAF AVLCAAWDAA TNKQIKPWSE DRELPELNDI 60

YNMNMSIASE EWQTIFDGQA EQQTWSQFAQ ANAGKYKGID WKQNWDRWRK QRQQTKDAGG 120

AWQTKNHRPE WAATPRDVRP VILAIAEEAT ELSRKLEPPR TADGKDLIAE INSKLASARC 180

SGELKAAAGN IGCTGPEGTP DKTTTCTTAK AGGSIGHDML CLCSVAEATD KCSSTGVGDA 240

VPNSGEKLRS NGFQHIVARC PKGPESGTLP QAIDLALAML ATALGTQQPG SNNMILGKSG 300

GGTCTATNSA CVDYHEKFSK QQAGITGIPW VALLQQARAL YGTYVDAKLA AQTARQQIVM 360

LAGQAKREYR RPAGSLKDPA GVIQEQATNR RRHGADDTNQ CTSNNATADE CPETRCEYDS 420

EKNECRPKKG TETTATGPGE RTTPADGKAN NTVSDSLLIK TSPLWLAFLL F

ILTat 1·22 (MIAG 202) EMBL X56765

MDTAQVFALF YMATVMAAGT KNKASQAVSD PCSEIHFDEQ LANYFENEVS AATTQLDENQ 60

NFERSWKLLQ YLQMDHQKSK GAAALAAYAS TINIRTAANV KAASGELLTA ASLLRQRAAN 120

VSAAFQLQGQ GVIKLGTPDI DNGAKSITHA DAGCNYAAIS KTVPTQRCTP PQQQADTITA 180

ADMQPDKLDE LQLITEAYTT TITIAASAYS KGTPATGHTV YTYGNCQSTG GSASAQLGDT 240

HALGIHVKTI GTKAVTEKTT LQPSSSNKCP DEGTTAELTP IKRLARAICL ARKASLAKPK 300

ALSRLQYSDL QTDTDFKRIA AIFLSRNGKQ LDPEKDSQEI NELIKETYGP NEEHFHKSYV 360

EALDNKKWEF KIKESKIEGT VNALANGVDA GLATAYYASK RQSTCGQAAA DTPIVSSDVE 420

KCKGKTQDDC RTADECEMRD GECNAKVAKT AEPDSKTNTT GNNSFAIKTS TLLLAVLLF

Fig. 1.

ILTat 1·23 (MIAG 206) EMBL X56768

MFKNINAAVL LLILSTRNDY ANAAAGDNEN LFLDLCNLLE LGKRAVSKLA PTNLGELAYS 60

EIQKLNMSLS DDAWKAKFAP KKGKQENSNN ADGAPRSQEK THARNHAWRQ AATDLAGDEG 120

DKPTLRLAGL EAVTQVEKLQ YLSALQPLAE RAAAILEQLK TLHSGSAGLT DTNIRQEIQT 180

ALYGTGATAP EKTTLQLLKG KGNVGSTRKD ICGQDNTAAK ADTVLAYLFC ICAGHATDSG 240

GAIKVCSQTQ PANNKADADV SDAHTHAAAL AGQCHGSDTT NDIKAAEIDS AILEFTSKLK 300

AANQKPYFGK YSATGCTGSD AEGICVMFKT TAKGEGKAVK QIPWVLTLHN AAEMIRKQQA 360

VNGKIDSLNQ ELQAIQTAAY ALKPQLEMYK RLQQTTEKAR PGKQLTEMQA GECNTHKSNS 420

TCPKNNCKWE EKDGKDGKCV ADDSKVTTQG NAPAGAGDGT AGTTTTPNCA SHTDKTKCEE 480

ENKGKTTPVC GWRKGKEGES DQDKEMCRNG SFLAKKKFAL SVVSAAFTAL LF

ILTat 1·24 (MIAG 209) EMBL X56767

MVYRNILQLS VLKVLLIVLI VEATHFGVKY ELWQPECELT AELRKTAGVA KMKVNSDLNS 60

FKTLELTKMK LLTFAAKFPE SKEALTLRAL EAALNTDLRA LRDNIANGID RAVRATAYAS 120

EAAGALFSGI QTLHDATDGT TYCLSASGQG SNGNAAMASQ GCKPLALPEL LTEDSYNTDV 180

ISDKGFPKIS PLTNAQGQGK SGECGLFQAA SGAQATNTGV QFSGGSRINL GLGAIVASAA 240

QQPTRPDLSD FSGTARNQAD TLYGKAHASI TELLQLAQGP KPGQTEVETM KLLAQKTAAL 300

DSIKFQLAAS TGKKTSDYKE DENLKTEYFG KTESNIEALW NKVKEEKVKG ADPEDPSKES 360

KISDLNTEEQ LQRVLDYYAV ATMLKLAKQA EDIAKLETEI ADQRGKSPEA ECNKITEEPK 420

CSEEKICSWH KEVKAGEKNC QFNSTKASKS GVPVTQTQTA GADTTAEKCK GKGEKDCKSP 480

DCKWEGGTCK DSSILANKQF ALSVASAAFV ALLF

ILTat 1·25 (MIAG 211) EMBL X56769

MQSQQQPVFI SIILLAINTD AAAPTTAVNA REFGLLCTLV RAEDNLEDRR SAGQAAKEVV 60

ALAAKIELIL ANLKHIERLA AAEPEAAPKE SRSDETPEAC KASKATVCSQ AAQTYKTIRP 120

DEKLALAFLA ETTGDLRSTF NVTLKQITAA AASHARYFGE NTESRPALDK IKKALYGSPE 180

AKGDAIIESG DGTRSAACGN TDGNAANSAA KRATAALICL CGGDNTNTGN DACFTQTTAD 240

INYAKKGGEV ERAWTEIRQK CKAGAAANKV TAAQLKAAAA ELAALIHQKR GEKAVAGLLG 300

AAQINNGAVD CDGSEANGKG SCVILSTSAS KYKVETPDWL NALEAAIADL EQEQIELDNG 360

RKAEAQILAL NSS$^{P}_{L}$TTLLAQ AVEPTKQPPA KAAAAPEKKS NPQKDCNKNT KKRDCKEGDG 420

CKWSSTEATE GAFCKPKDGE GQTSAAGAGD AGASDTEAKK CSDKKKEEEC KSPNCKWDGK 480

ECKDSSILAN KQFALSVASA AFVALLF

Fig. 1.

isolated from libraries; ILTat 1·21 and ILTat 1·22, partial sequence from cDNAs isolated from libraries and complete sequence from PCR generated cDNAs; MITat 1·1, MITat 1·5, MITat 1·6 and ILTat 1·23, PCR generated cDNAs. The ambiguity at residue 374 of ILTat 1·25 VSG is due to a base difference in the two library cDNA clones sequenced.

(b) *Primary structure of the VSG polypeptides*

The amino acid sequences of VSGs MITat 1·1, MITat 1·2, MITat 1·5, MITat 1·6, ILTat 1·21, ILTat 1·22, ILTat 1·23, ILTat 1·24 and ILTat 1·25 are presented in Figure 1. The locations of cleavage of the N-terminal signal sequence were determined for each VSG by comparison of the cDNA-derived sequence with that determined by direct N-terminal sequencing of purified VSG polypeptide (ILTat 1·21, Gurnett, 1983; MITat 1·1 and MITat 1·5, P. Voorheis, personal communication; the remainder, Metcalf *et al.*, 1988). The N-terminal signal sequences of all the VSGs are typical of those found on eukaryotic polypeptides co-translationally secreted into the endoplasmic reticulum and, as previously reported (Cross, 1984), do not show any great homology to one another. The C termini of the mature VSG polypeptides from the MITAR 1 serodeme were located by comparison of the cDNA-derived sequence with the directly determined C-terminal sequence (Holder & Cross, 1981). The designation of the mature C terminus of the ILTAR 1 VSGs was based on sequence homology with members of the MITAR 1 serodeme.

(c) *Sequence homology between VSGs, conservation of cysteine residues in the N-terminal domain*

The amino acid sequences of the nine VSG N-terminal domains (nominally taken to be residues 1 to 350 of the mature VSG polypeptide) were compared with the University of Wisconsin Genetics Computing Group GAP programme which estimates pecentage identity after the introduction of gaps to maximize homology (Devereux *et al.*, 1984). Of the comparisons 34/36 gave values below 25% identity, MITat 1·1 and MITat 1·2 have 27.6% identity and MITat 1·2 and ILTat 1·24 have 26·6% identity. Since random polypeptides aligned in this manner give values of 10 to 20% identity (Doolittle, 1986) and the nine VSGs have similar amino acid compositions these values for percentage indentity indicate that there is probably little homology between the N-terminal domains of the VSGs presented here and so they represent nine new examples of functional VSGs.

The location of cysteine residues in the N-terminal domain divides the VSGs into three groups based on the relative positions of these residues (Fig 2(a)). All the VSGs have a conserved cysteine residue between positions 13 and 17 from the mature N terminus, however, other than this cysteine each of the three groups appears to have a distinct pattern of conserved cysteine residues. The majority of VSGs, including all the members of the MITAR 1 serodeme sequenced thus far, are type A and are characterized by the four conserved cysteine residues. Of the eleven type-A VSGs, six have just these four cysteine residues in their N-terminal domains (MITat 1·1, MITat 1·2, ILTat 1·3, ILTat 1·24, ANTat 1·1 and ANTat 1·10). Three of the others have an extra pair of cysteine residues the position of which is variable (MITat 1·5, MITat 1·6 and ILTat 1·22), MITat 1·4 has one and MVAT 4 has three additional cysteine residues.

There are three N-terminal domain type-B *T. brucei* VSGs (ILTat 1·21, ILTat 1·23 and ILTat 1·25); these VSGs have eight conserved cysteine residues. ILTat 1·21 and ILTat 1·25 each have an additional pair of cysteine residues, the position of which is not conserved. Also shown in Figure 2 are two VSGs (YNat 1·1 and YNat 1·3) from *Trypanosoma congolense* (Strickler *et al.*, 1987) which are truncated when compared with VSGs from *T. brucei*, as they have no C-terminal domain. In addition to the eight conserved cysteine residues, an additional pair occurs in both ILTat 1·21, ILTat 1·25 and YNat 1·1; and a single additional cysteine residue occurs in YNat 1·3. None of these residues occurs in a position that is conserved between different VSGs.

There are two *T. brucei* VSGs that have been classed as N-terminal domain type C (ILTat 1·1 and MVAT 7) as has VSG 20 from *T. equiperdum* (Thon *et al.*, 1990). These have six conserved cysteine residues; ILTat 1·1 has an additional two cysteine residues not found in MVAT 7.

The potential *N*-linked glycosylation sites (Asn–X–Ser/Thr) are shown in Figure 2. Generally there appears to be little conservation in the position of these sites in the primary structure of the N-terminal domains.

(d) *Sequence homology between VSGs, conservation of cysteine residues and glycosylation sites in the C-terminal domain*

VSGs have previously been divided into two groups based on the number of cysteine residues in the C-terminal domain; group 1 VSGs have eight cysteine residues and group 2 have four or five (Rice-Ficht *et al.*, 1981, Turner, 1988). Figure 2(b) represents the primary structure of the VSGs aligned on their mature C termini showing the location of cysteine residues and potential *N*-linked glycosylation sites. With the additional data available the VSGs have been grouped into four proposed C-terminal types. These groupings are based on: (1) the number of and location, relative to the C terminus, of cysteine residues; (2) the identity of the C-terminal residue; (3) the sequence of the C-terminal hydrophobic extension and (4) the position of potential glycosylation sites. Figure 3 shows the amino acid sequences of the C-termini and hydrophobic extensions of these VSGs.

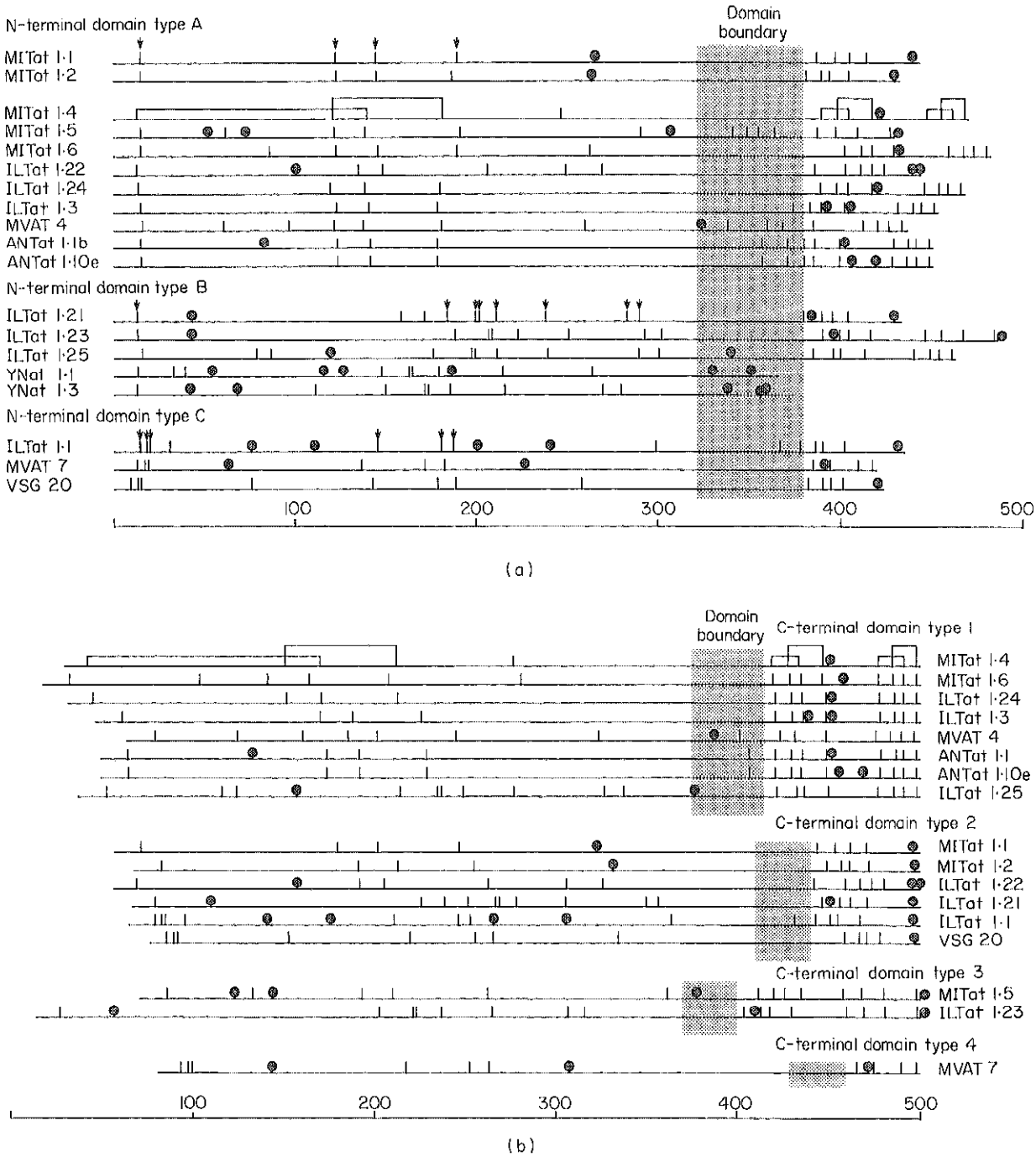

Figure 2. Representation of the primary structures of VSGs for which the complete amino acid sequence is known. The mature polypeptide is shown, the location of cysteine residues are marked by vertical bars and potential *N*-linked glycosylation sites by the filled circles. Disulphide linkages are shown on MITat 1·4, the only VSG for which these have been determined. The position of the domain boundary is approximate and is based on the size of fragments produced by mild proteolytic digestion (Johnson & Cross, 1979, Metcalf *et al.*, 1988) and on the results of the crystal structure solved for MITat 1·2 (Freymann *et al.*, 1990) and ILTat 1·24 (Blum *et al.*, 1990). (a) The sequences aligned at their N termini to show grouping into N-terminal domain types. The arrows above the 1st member of each group show the cysteine residues conserved within that group. (b) The sequences aligned at their C termini to show grouping into C-terminal domain types.

Eight VSGs have type 1 C-terminal domains containing eight cysteine residues in two groups of four. Each of these groups of four forms two internal disulphide bonds in MITat 1·4 (Allen & Gurnett, 1983) and it is probable that the other members of the group form the same disuphide linkages. All the VSGs in this group have aspartic acid as the C-terminal residue of the mature polypeptide and have a hydrophobic extension of 23 residues that starts serine–serine–. Type-1 C-

C-terminal domain type 1

```
                    **  *     *     ***    ***   ** *  * * *   ****  ***
MITat 1·4    EATTDKCKGKL-EDTCKKESNCKWENNACKD   SSILVTKKFALTVVSAAFVALLF
MITat 1·6    EATPEKCKGKDAK-TCGTTQGCKWEGETCKD   SSILVTKKFALTVVSAAFVALLF
ILTat 1·24   DTTAEKCKGKGEKD-C-KSPDCKWEGGTCKD   SSILANKQFALSVASAAFVALLF
ILTat 1·3    QTTTDKCKDKKK-DDC-KSPDCKWEGETCKD   SSFILNKQFALSVVSAAFAALLF
MVAT 4       DKKEEKCKGKLE-PECTKAPECKWEGETCKD   SSILVNKQFTLSMISAAFM...
ANTAT 1·1    ETPAEKCTGKKK DDC-K-DGCKWEAETCKD   SSILVTKKFALSLVSAAFASLLF
ANTAT 1·10e EATTDKCKDKTK-DEC-KSPNCKWEGETCKD   SSILVTKKFALSLVSAAFASLLF
ILTat 1·25   EA--KKCSDKKKEEEC-KSPNCKWDGKECKD   SSILANKQFALSVASAAFVALLF
```

C-terminal domain type 2

```
                                  * *    * *      *  * *   * ** * *
MITat 1·1    LKKDVKEKLEKESKETEGKDEK---ANTTGS   NSFLIHKA-PLLLAFLAF
MITat 1·2    KKCTLDKEEAKKVADETAKDGKTGNTNTTGSS  NSFVISKT-PLWLAVLLF
ILTat 1·22   ECEMRDGECNAKVAKTAEPDSK---TNTTGN   NSFAI-KTSTLLLAVLLF
ILTat 1·21   PKKGTETTATGPGERTTPADGK---ANNTVS   DSLLI-KTSPLWLAFLLF
ILTat 1·1    EEAKRVAEQAATNQETEGKDGKT--TNTTGS   NSFVIHKA-PLFLAFLLF
```

C-terminal domain type 3

```
              **      *** **** ** *  *** ***   ***   * *  ******* ***
MITat 1·5    KENTPGQSAVCGFRKGKDGETDEPDKEKCRN   GSFLTSKQFAFSVVSAAFMALLF
ILTat 1·23   EENKGKTTPVCGWRKGKEGESDQ-DKEMCRN   GSFLAKKKFALSVVSAAFTALLF
```

C-terminal domain type 4

```
MVAT7        IQNQTECEGVKGTPPTGKAKVCGWIEGKCQD   SSFLLSKQFALSVVSAAFAALLF
```

Figure 3. The amino acid sequences of the C termini and hydrophobic extensions of VSGs. The gap in the sequence represents the position of the cleavage to form the mature C terminus. The location of the mature C terminus is based on sequence homology to members of the MITAR 1 serodeme for which the identity of the mature C terminus has been determined (Holder & Cross 1981). Some gaps (–) have been introduced to increase sequence homology.

terminal domains have a glycosylation site either between the two groups of four cysteine residues (*MITat 1·4, MITat 1·6, ILTat 1·24, ILTat 1·3,* ANTat 1·1 and ANTat 1·10) or in the region of the boundary between the N and C-terminal domains (MVAT 4 and ILTat 1·25). ANTat 1·1 and ANTat 1·10 have an additional cysteine residue in the region of the domain boundary and MVAT 4 has apparently lost the sixth cysteine residue from the C terminus but has an extra cysteine residue in the domain boundary region; it is not clear from this whether MVAT 4 represents a variant of the type-1 C-terminal domain or a novel structure.

There are five VSGs with a type 2 C-terminal domain. These all have a group of four cysteine residues followed by between 19 and 33 residues before the mature C terminus. The C-terminal residue is serine, except for ILTat 1·22 which has a C-terminal asparagine residue that is also a potential glycosylation site. The hydrophobic extension is 17 residues long and usually starts with asparagine although ILTat 1·21 starts with aspartic acid. All have a glycosylation site four to six residues from the C terminus. The position of cysteine residues relative to the C terminus is more variable than in the type-1 VSGs as is the sequence of the hydrophobic extension. ILTat 1·1 and ILTat 1·22 have an additional cysteine residue towards the region of the boundary between the N and C-terminal domains.

The type-3 C-terminal domain (MITat 1·5 and ILTat 1·23) has some similarities to type 1 in that the domain contains eight cysteine residues. Also the group of four cysteine residues furthest from the C terminus in types 1 and 3 is similar in terms of spacing between cysteine residues. The difference between types 1 and 3 lies in the four cysteine residues closest to the C terminus. The relative positions of the four cysteine residues nearest the C terminus is quite different and the C-terminal residue in type 3 is asparagine, which is also a potential glycosylation site. The hydrophobic extension is 23 residues long and starts glycine–serine–.

The C-terminal domain of MVAT 7 is put in a class of its own, type 4. It contains four cysteine residues as also does type 2. However, it differs in the following ways: the cysteine nearest to the C terminus is only three residues distant; the C-terminal residue is aspartic acid and the hydrophobic extension is 23 residues in length. There is a glycosylation site 29 residues from the C terminus in MVAT 7 compared with a site four to six residues from the C terminus found in type 2 VSGs.

(c) *Classification of VSGs as a combination of domain types*

The combination of N-terminal and C-terminal domain found in each VSG is as follows: MITat 1·1, A2; MITat 1·2, A2; MITat 1·5, A3; MITat 1·6, A1; ILTat 1·21, B2; ILTat 1·22, A2; ILTat 1·23, B3; ILTat 1·24, A1 and ILTat 1·25, B1. It is possible to add to this list the domain combinations of previously published VSG cDNAs, MITat 1·4, A1; ILTat 1·1, C2; ILTat 1·3, A1; MVAT4, A1; MVAT7, C4; ANTat 1·1, A1 and ANTat 1·10e, A1. It is immediately apparent that a range of combinations are found. This range is probably only limited by the amount of data available.

4. Discussion

(a) *Domain combinations and structure in VSGs*

Comparison of known VSG sequences, the new sequences presented here along with previously published sequences, reveals that, in both the N-terminal and C-terminal domain, the VSG domains can be classified based on the pattern of conserved cysteine residues (Fig. 2). The disulphide bonds formed by the cysteine residues in the N-terminal domain have been determined chemically for MITat 1·4 (Allen *et al.*, 1982) and crystallographically for MITat 1·2 (Freymann *et al.*, 1990) and are conserved. The strict geometry constraints necessary to form a disulphide bond imply that conserved disulphide linkages indicate a conserved secondary and tertiary structure (Richardson & Richardson, 1989). The disulphide bond between Cys15 and Cys145 in MITat 1·2 covalently bridges a surface loop to one of the core helices of the N-terminal domain structure. The second disulphide bond ties a strand passing over the top of the domain to a buried part of the chain (Fig. 5 in Freymann *et al.*, 1990). These features can be expected to be present in the other VSG variable domains in which these cysteine residues are conserved, and has been confirmed in a second class-A N-terminal domain (Blum, 1990; M. Blum *et al.*, unpublished results). The other sequence classes have a larger number of conserved cysteine residues. It is not yet possible to rule out that the type-B and type-C N-terminal domain cysteine residues are a superset of the class-A cysteine residues. This is implied by the strict conservation, across all N-terminal classes, of the cysteine residue near position 15, but multiple sequence alignment (Subbiah & Harrison, 1989) using both class A and B sequences (M. Blum, unpublished results) has not established the relationship of the B domains to the class-A structure, nor identified the corresponding second cysteine residue of the expected disulphide bond in the class B sequences.

Is there is any structural limitation on an individual N-terminal domain limiting the type of C-terminal domain found in the VSG? Could a genetic recombination occur so that the N-terminal domain found in MITat 1·1 (A 2) formed a functional VSG with an A 1 or A 3 domain combination? Genetic recombination is known to occur around the region of the gene encoding the domain boundary, one example occurred between two antigenically distinct members of the same VSG gene family (ANTat 1·1 and ANTat 1·10) with the same domain types (Pays *et al.*, 1983). A second example is that the C-terminal 80 residues of ILTat 1·1 and *T. equiperdum* VSG 78 (Roth *et al.*, 1986) are of

C-terminal domain type 1		*§ * * # §*		§*§ *§ *§# §*§§	
MITat 1·4		ECN 6 KCN 3 ICSW 9 HCK	27	KCK 5 TCK 3 NCKW 4 ACKD	
MITat 1·6		SCN 6 KCN 3 KCSY 8 KCK	27	KCK 5 TCG 3 GCKW 3 TCKD	
ILTat 1·24		ECN 6 KCS 3 ICSW 9 NCQ	26	KCK 5 DCK 2 DCKW 3 TCKD	
ILTat 1·3		TCN 6 ACN 3 FCTY 9 KCK	26	KCK 5 DCK 2 DCKW 3 TCKD	
MVAT 4	ACK 19 RCE 5 ACP 14 ECK	24		KCK 5 ECT 3 ECKW 3 TCKD	
ANTAT 1·1	ACP 11	LCS 6 ECN 3 FCSY 9 KCK	27	KCT 5 DCK 1 GCKW 3 TCKD	
ANTat 1·10e	PCP 11	LCS 6 ECN 3 FCSY 9 KCQ	26	KCK 5 ECK 2 NCKW 3 TCKD	
ILTat 1·25		DCN 6 DCK 3 GCKW 9 FCK	24	KCS 6 ECK 2 NCKW 3 ECKD	
C-terminal domain type 3					* *§ §*§# §*§§
MITat 1·5		NCD 5 ECG 3 GCGW 5 KCE		20	RCT 7 ACE 9 VCGF 14 KCRN
ILTat 1·23		ECN 6 TCP 2 NCKW 8 KCV		27	NCA 6 KCE 9 VCGW 13 MCRN
C-terminal domain type 4					
MVAT7					FCS 6 ECE 12 VCGW 3 KCQD
C-terminal domain type 2					
MITat 1·1		ICN 7 ACE 5 GCVF 5 KCE	28		
MITat 1·2		SCN 5 ECK 1 PCKW 7 KCT	28		
ILTat 1·22	TCG	14		KCK 5 DCR 3 ECEM 3 ECN 18	
ILTat 1·4		ECK 5 TCK 1 GCKW 7 KCV	25		
ILTat 1·21		QCT 7 ECP 2 RCEY 5 ECR	28		
ILTat 1·1	ECN 8	TCE 6 NCK 1 PCKE 9 KCK	32		
		α		β	γ

Figure 4. The relative positions of cysteine residues and conserved flanking residues in the C-terminal domain of VSGs. The standard 1-letter amino acid code is used, the numbers represent the size of the gaps in numbers of amino acids not shown. The symbols above the residues indicate (*) cysteine residues; (§) conserved residue in >60% of examples; (#) conserved aromatic amino acid.

common origin (54/80 identities with gaps introduced), at 80 residues from the C terminus there is an abrupt end to the homology so it is probable that the two VSGs have different N-terminal domains. However, the distribution of glycosylation sites suggests some limitations in that the type-A N-terminal domains have a potential *N*-glycosylation site in all examples of VSGs of domain combinations A2 ($n=3$) and A3 ($n=1$) but only one (ANTat 1·1) from seven examples of A1 VSGs has such a site (Fig. 2). This suggests a difference between the N-terminal domains capable of forming a functional VSG in combination with a type 1, as opposed to types 2 and 3, C-terminal domain.

There are more obvious similarities between the C-terminal domain types. The disulphide linkages are known for the C-terminal domain of one type-1 VSG, MITat 1·4 (Fig. 2(b)). The eight cysteine residues can be divided into two groups of four, each of which shows the same disulphide bond formation with respect to position of the constituent cysteine residues in the primary structure. It has been suggested on the grounds of protease resistance that each of these groups of four cysteine residues represents a tightly folded structural domain (Allen & Gurnett, 1983). Since each of the C-terminal domain types is based on either four or eight conserved cysteine residues and each group of four probably represents a small domain, VSGs can be visualised in *T. brucei* as a variable N-terminal domain with one (types 2 and 4) or two (types 1 and 3) smaller and relatively conserved domains at the C terminus.

It has been suggested that the type-2 C-terminal domains, which contain four cysteine residues, arose from the type 1 through a deletion event (Cross, 1984). By considering the spacing between the cysteine residues, and the identity of the residues flanking them it is possible to suggest that the groups of four cysteine residues are divided into three patterns, α, β and γ (Fig. 3). The different C-terminal domain types are derived from unequal crossing overs at the DNA level generating combinations of these groups, type 1, $\alpha\beta$; type 3, $\alpha\gamma$; type 2, α except ILTat 1·22, β; type 4, γ. The exception is MVAT 4 which nominally has a type 1 C-terminal domain; however, the first set of four cysteine residues does not match any of the α, β or γ groupings.

It can also be seen in Figure 4 that the identity of the amino acid, two residues to the C-terminal side of the third cysteine in each group of four, is one of the aromatic amino acids (tryptophan, tyrosine or phenylalanine) in 22/24 of the examples shown. It may be significant that the two VSGs that form exceptions to this rule (ILTat 1·22 and ILTat 1·1) are the only two VSGs with a type 2 C-terminal domain with a fifth cysteine residue. The significance of this extra cysteine is clear, in a test of seven VSGs (ILTats 1·3, 1·4, 1·21, 1·22, 1·23, 1·24, 1·25) it was found that only ILTat 1·22 occurred as a disulphide linked dimer in solution (Miller, 1980). Disulphide linked VSG dimers have also been reported to occur in *T. evansi*. (Cross, 1977). A prediction from this result is that VSG ILTat 1·1 may occur as a disulphide linked dimer in solution. The same func-

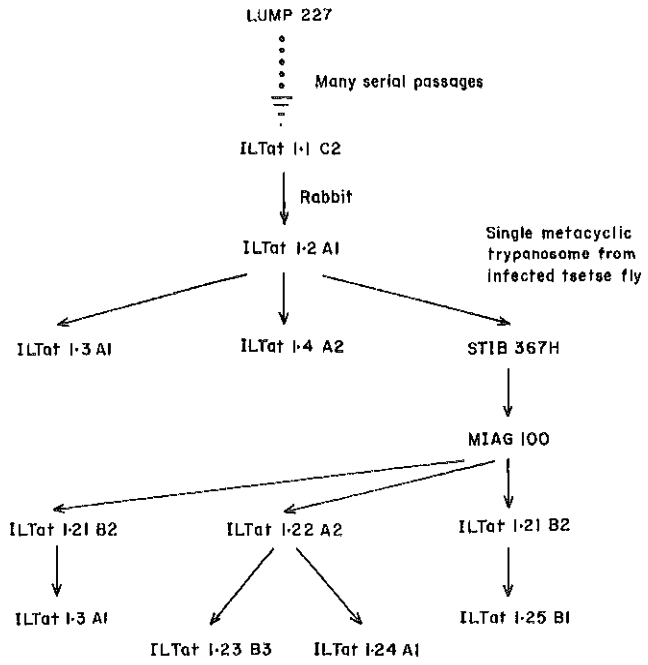

Figure 5. Simplified derivation of ILTAR 1 serodeme showing the VSGs expressed by cloned populations and the domain combinations present in those VSGs. See Miller & Turner (1981) for full details of the serodeme. All passages were in mice unless shown otherwise on this Figure. LUMP 227 is derived from isolate EATRO 795.

tion cannot be attributed to the odd cysteine residues in N-terminal domains as MITat 1·4 does not occur as a disulphide linked dimer (Allen & Gurnett, 1983).

(b) *Domain combinations present in each repertoire*

There are five characterized VSGs from the MITAR 1 repertoire; all have a type-A N-terminal domain, whereas in the seven characterized members of the ILTAR 1 serodeme, N-terminal domain types A, B and C are found. It could just be chance that all the MITAR 1 VSGs have the same N-terminal domain type; however, there are two other possibilities. Either the MITAR 1 serodeme could be genetically more restricted, or the method of generating the cloned trypanosome populations for each repertoire may have had some effect. The MITAR 1 serodeme was generated by cloning trypanosomes in mice after isolation from a rabbit at different time points after infection, whereas ILTat 1·21 to ILTat 1·25 were generated using mice only.

(c) *Biological context of VSG domain combinations*

The best characterized isolate of *T. brucei* in terms of sequence information about VSGs is the ILTAR 1 serodeme. Seven of nine VSGs described are sequenced fully and there are sufficient sequence data for the other two to designate their domain combinations with reasonable certainty, ILTat 1·2, A1(Rice-Ficht *et al.*, 1981); ILTat 1·4, A2 (Rice-Ficht *et al.*, 1981; Donelson *et al.*, 1982). The domain combinations can be superimposed on the relationship between the cloned trypanosome stocks from the ILTAR 1 serodeme (Miller & Turner, 1981; Fig. 5). It can be seen in all cases except one that on switching of VSGs, the new VSG has a different domain combination to the old VSG. At first glance, this provides an attractive model for antigenic variation as it could allow the trypanosome a mechanism to distinguish "old" and "new" VSGs. However, this switching from one domain combination to another is not too surprising as among the nine ILTAR 1 VSGs six different domain combinations are found. Furthermore, there are two clear examples where no such switch occurs, ILTat 1·2 to ILTat 1·3 in these experiments and ANTat 1·1 to ANTat 1·10 (Pays *et al.*, 1983). The above changes in VSG expression were analysed by cloning individual trypanosomes from the population before and after switching. An approach that samples more switching events is needed. Such data are available from more general experiments in which approximately ten trypanosomes from a cloned population were used to infect mice and variants arising in the second peak of parasitaemia were recognized using antisera raised against known variant types (Miller & Turner, 1981). In this way, the percentage of the population corresponding to different domain combinations can be estimated to some extent. The most informative experiments are perhaps those in which trypanosomes expressing ILTat 1·22 switched to a range of VSGs, but the predominant switch was to ILTat 1·4. Both ILTat 1·22 and ILTat 1·4 are A 2 VSGs, again confounding the hypothesis.

Do the observations above indicate that the domain combination of VSGs has nothing to do with the process of antigenic variation but is merely a result of structural constraints needed to produce a functional VSG? One observation that indicates that there may be more than structural considerations involved is in the identity of VSGs on a dual expressor (that is, a trypanosome expressing more than one VSG at a time). Baltz *et al.* (1986) described a dual expressing population which arose in culture. However the two *T. equiperdum* VSGs involved are of unknown domain combinations. Some earlier experiments with *T. brucei* (Miller, 1980) involved growing a cloned population expressing ILTat 1·3 in culture; the trypanosomes were grown for 28 days, then the culture was neutralized with anti-ILTat 1·3 serum and used to infect a mouse. After four days the VSGs being expressed were characterized by immunofluorescence. All the trypanosomes appeared to be expressing both ILTat 1·4 and ILTat 1·22 in roughly equivalent amounts. This population was used to serially infect a second mouse; after a further two days all the trypanosomes were still positive with both antisera but the reaction with anti-ILTat 1·4 was much stronger than the reaction with anti-ILTat 1·22. Both ILTat 1·4 and ILTat

1·22 are A 2 VSGs, the only two A 2 VSGs from the nine characterized VSGs of the ILTAR 1 serodeme, so the likelihood that the dual expressor has two VSGs of the same domain combination by chance is small. These results suggest that a period in culture occasionally leads to expression of two VSGs by one trypanosome and that if the two VSGs are of the same domain combination then the dual expression is stable for a few days when transferred to an animal. Trypanosomes expressing two VSGs of different domain combinations are either unstable in culture or in the animal, and rapidly switch to a trypanosome expressing one or the other but not both. Under the normal *in vivo* conditions for antigenic variation, a rapid switch is desirable. Therefore variants in which a switch in domain type has occurred would be at an advantage over those in which the domain combination remained the same.

We thank Linda Allen and Paul Vorheis for help in preparing bloodstream-form trypanosomes. This work was funded by Merck Sharp & Dohme and the Wellcome Trust. The oligonucleotide synthesis facility in the Biochemistry Department in Cambridge is supported by the Wellcome Trust.

References

Allen, G., Gurnett, L. P. & Cross, G. A. M. (1982). Complete amino acid sequence of a variant surface glycoprotein (VSG 117) from *Trypanosoma brucei*. *J. Mol. Biol.* **157**, 527–546.

Allen, G. & Gurnett, L. P. (1983). Locations of the six disulphide bonds in a variant surface glycoprotein (VSG 117) of *Trypanosoma brucei*. *Biochem. J.* **209**, 481–487.

Auffret, C. A. & Turner, M. J. (1981). Variant specific antigens of *Trypanosoma brucei* exist in solution as glycoprotein dimers. *Biochem. J.* **193**, 647–650.

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Séguin, C., Tufnell, P. S. & Barrell, B. G. (1984). DNA sequence and expression of the B95-8 Epstein–Barr virus. *Nature (London)*, **310**, 207–211.

Baltz, T., Giroud, C., Baltz, D., Roth, C., Raibaud, A. & Eisen, H. (1986). Stable expression of two variable surface glycoproteins by cloned *Trypanosoma equiperdum*. *Nature (London)*, **319**, 602–604.

Blum, M. (1990). The structure of a VSG variable domain from *Trypanosoma brucei* at 2·7 Å resolution. Ph.D. thesis, Harvard University.

Boothroyd, J. C., Paynter, C. A., Cross, G. A. M., Bernards, A. & Borst, P. (1981). Variant surface glycoproteins of *Trypanosoma brucei* are synthesized with cleavable hydrophobic sequences at the carboxy and amino termini. *Nucl. Acids Res.* **9**, 4735–4743.

Boothroyd, J. C., Paynter, C. A., Coleman, S. L. & Cross, G. A. M. (1982). Complete nucleotide sequence of complementary DNA coding for a variant surface glycoprotein from *Trypanosoma brucei*. *J. Mol. Biol.* **157**, 547–556.

Borst, P. (1986). Discontinous transcription and antigenic variation in trypanosomes. *Annu. Rev. Biochem.* **55**, 701–732.

Capbern, A., Giroud, C., Baltz, T. & Mattern, P. (1977). Etude des variation antigenique au cours de la trypanosomose experimentale du lapin. *Expl. Parasit.* **42**, 6–13.

Cardoso de Almeida, M. L. & Turner, M. J. (1983). The membrane form of variant surface glycoproteins of *Trypanosoma brucei*. *Nature (London)*, **302**, 349–353.

Cornelissen, A. W. C. A., Vespieren, M. P., Toulmé, J.-J., Swinkels, B. W. & Borst, P. (1986). The common 5' terminal sequence on typanosome mRNAs, a target for anti-messenger oligodeoxynucleotides. *Nucl. Acids Res.*, **14**, 5605–5615.

Cross, G. A. M. (1975). Identification, purification and properties of clone specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. *Parasitology*, **71**, 393–417.

Cross, G. A. M. (1977). Isolation, structure and function of variant specific surface antigens. *Annu. Soc. Belge Med. Trop.* **57**, 389–399.

Cross, G. A. M. (1984). Structure of the variant glycoproteins and surface coat of *Trypanosoma brucei*. *Phil. Trans. Roy. Soc. London*, **B307**, 3–12.

Cross, G. A. M. & Manning, J. C. (1973). Cultivation of *Trypanosoma brucei* spp. in semi-defined and defined media. *Parasitology*, **67**, 315–331.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **112**, 387–395.

Donelson, J. E., Young, J. R., Dorfman, D., Majiwa, P. A. O. & Williams, R. O. (1982). The ILTat 1.4 surface antigen gene family of *Trypanosoma brucei*. *Nucl. Acids Res.* **10**, 6581–6595.

Doolittle, R. F. (1986). *Of URFS and ORFS*. Oxford University Press.

Englund, P. T., Hereld, D., Krakow, J. L., Doering, T. L., Masterson, W. J. & Hart, G. W. (1988). Glycolipid membrane anchor of the trypanosome variant surface glycoprotein, its biosynthesis and cleavage. In *The Biology of Parasitism* (Englund, P. T. & Sher, A., ed.), pp. 401–412, Alan R. Liss, New York.

Ferguson, M. A. J., Homans, S. W., Dwek, R. A. & Rademacher, T. W. (1988). Glycosylphosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane. *Science*, **239**, 753–759.

Freymann, D., Metcalf, P., Turner, M. J. & Wiley, D. C. (1984). 6 Å resolution X-ray structure of a variable surface glycoprotein from *Trypanosoma brucei*. *Nature (London)*, **311**, 167–169.

Freymann, D., Down, G., Carrington, M., Roditi, I., Turner, M. J. & Wiley, D. C. (1990). 2·9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from *Trypanosoma brucei*. *J. Mol. Biol.* **216**, 141–160.

Gurnett, A. M. (1983) Structure of variant surface glycoproteins from *Trypanosoma b brucei*. Ph.D. thesis, University of Cambridge.

Holder, A. A. & Cross, G. A. M. (1981). Glycopeptides from variant surface glycoproteins of *Trypanosoma brucei*. C-terminal location of antigenically cross-reacting carbohydrate moieties. *Mol. Biochem. Parasitol.* **2**, 135–150.

Johnson, J. G. & Cross, G. A. M. (1979). Selective cleavage of variant surface glycoproteins from *Trypanosoma brucei*. *Biochem. J.* **178**, 689–697.

Kawasaki, E. S. (1990). Amplification of RNA. In *PCR Protocols* (Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J., eds), pp. 21–27, Academic Press, London.

Leonardo, M. J., Rice-Ficht, A. C., Kelly, G., Esser, K. M. & Donelson, J. E. (1984). Characterization of the genes specifying two metacyclic variable antigen types in *Trypanosoma brucei rhodesiense*. *Proc. Nat. Acad. Sci., U.S.A.* **81**, 6642–6646.

Majumder, H. K., Boothroyd, J. C. & Weber, H. (1981). Homologous 3′-terminal regions of mRNAs for surface antigens of different antigenic variants of *Trypanosoma brucei*. *Nucl. Acids Res.* **9**, 4745–4753.

Matthyssens, G., Michiels, F., Hamers, R., Pays, E. & Steinert, M. (1981). Two variant specific glycoproteins have a conserved C-terminus. *Nature (London)*, **293**, 230–233.

Metcalf, P., Blum, M., Freymann, D., Turner, M. J. & Wiley, D. C. (1987). Two variant surface glycoproteins of *Trypanosoma brucei* of different sequence classes have similar 6 Å resolution X-ray structure. *Nature (London)*, **325**, 84–86.

Metcalf, P., Down, J. A., Turner, M. J. & Wiley, D. C. (1988) Crystallization of amino-terminal domains and domain fragments of variant surface glycoproteins from *Trypanosoma brucei*. *J. Biol. Chem.* **263**, 17030–17033.

Michiels, F., Matthyssens, G., Kronenberger, P., Pays, E., Dero, B., Van Assel, S., Darville, M., Cravador, A., Steinert, M. & Hamers, R. (1983). Gene activation and re-expression of a *Trypanosoma brucei* variant surface glycoprotein. *EMBO J.* **2**, 1185–1192.

Miller, E. N. (1980). Biological aspects of antigenic variation in trypanosomes. Ph.D. Thesis, University of Cambridge.

Miller, E. N. & Turner, M. J. (1981). Analysis of antigenic types appearing in first relapse populations of clones of *Trypanosoma brucei*. *Parasitology*, **82**, 63–80.

Pays, E., Delauw, M.-F., Van Assel, S., Laurent, M., Vervoort, T., Van Meirvenne, N. & Steinert, M. (1983). Modifications of a *Trypanosoma b brucei* antigen gene repertoire by different DNA recombinational mechanisms. *Cell*, **35**, 721–731.

Pays, E. & Steinert, M. (1988). Control of antigen gene expression in African Trypanosomes. *Annu. Rev. Genet*, **22**, 107–126.

Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. (1981). Sequence homologies near the C termini of the variable surface glycoproteins of *Trypanosoma brucei*. *Nature (London)*, **294**, 53–57.

Rice-Ficht, A. C., Chen, K. K. & Donelson, J. E. (1982). Point mutations during generation of expression-linked copy of trypanosome surface glycoprotein gene. *Nature (London)*, **298**, 676–679.

Richardson, J. & Richardson, D. C. (1989). Principles and patterns of protein conformation. *Prediction of protein structure and the principles of protein conformation*. (Fasman, G. D., ed.), Plenum Press, New York.

Roditi, I., Carrington, M., & Turner, M. (1987). Expression of a polypeptide containing a dipeptide repeat is confined to the insect stage of *Trypanosoma brucei*. *Nature (London)*. **325**, 272–274.

Roth, C., Longacre, S., Raibaud, A., Baltz, T. & Eisen, H. (1986). The use of incomplete genes for the construction of a *Trypanosoma equiperdum* variant surface glycoprotein gene. *EMBO J.* **5**, 1065–1070.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.

Strickler, J. E. & Patton, C. L. (1982). *Trypanosoma brucei*: nearest neighbor analyses on the major variable surface coat glycoprotein-crosslinking patterns with intact cells. *Exp. Parasitol.* **53**, 117–131.

Strickler, J. E., Binder, D. A., L'Italien, J. J., Shimamoto, G. T., Wait, S. W., Dalheim, L. J., Novotny, J., Radding, J. A., Konigsberg, W. H., Armstrong, M. Y. K., Richards, F. F. & Lalor, T. M. (1987). *Trypanosoma congolense*: structure and molecular organisation of the surface glycoproteins of two early bloodstream variants. *Biochemistry*, **26**, 796–805.

Subbiah, S. & Harrison, S. C. (1989). A method for multiple sequence alignment with gaps. *J. Mol. Biol.* **209**, 539–548.

Thon, G., Baltz, T., Giroud, C. & Eisen, H. (1990). Trypanosome varible surface glycoproteins: composite genes and order of expression. *Genes Develop.* **4**, 1374–1383.

Turner, M. J. (1988). Trypanosome variant surface glycoprotein. In *The Biology of Parasitism* (Englund, P. T. & Sher, A., eds), pp. 349–369, Alan R. Liss, New York.

Van der Ploeg, L. H. T., Valerio, D., De Lange, T., Bernards, A., Borst, P. & Grosveld, F. G. (1982). An analysis of cosmid clones of nuclear DNA from *Trypanosoma brucei* shows that the genes for VSGs are clustered in the genome. *Nucl. Acids Res.* **10**, 5905–5923.

Vickerman, K. (1969). On the surface coat and flagellar adhesion in trypanosomes. *J. Cell Sci.* **5**, 163–193.

Edited by J. Karn